The Role of Mitochondria in the Development of Insulin Resistance and Type 2 Diabetes

Direk, Kenan

Awarding institution:
King's College London

Title: The Role of Mitochondria in the Development of Insulin Resistance and Type 2 Diabetes

Author: Kenan Direk

The Role of Mitochondria in the Development of Insulin Resistance and Type 2 Diabetes

Kenan Direk BSc MSc

Thesis submitted for the degree of Doctor of Philosophy

King's College London

Supervised by Dr Toby Andrew and Dr Mario Falchi

2013

Table of Contents

Contribution of Collaborators

Chapter 3

Acquisition of computed tomography data: Dr Marina Cecelja.

Acquisition of all other data on the twin sample: Department of Twin Research at King's College London.

Variance inflation factor and residual analyses: Dr Toby Andrew.

Chapter 4

Gene expression data: Multiple Tissue Human Expression Resource.

Linkage disequilibrium maps: Dr Winston Lau and Dr Nikolas Maniatis.

Malécot test of association: Dr Winston Lau and Dr Nikolas Maniatis.

Chapter 5

Linkage disequilibrium maps: Dr Winston Lau and Dr Nikolas Maniatis.

Malécot test of association: Dr Winston Lau and Dr Nikolas Maniatis.

Acknowledgements

Firstly to my first supervisor Toby Andrew, who has transformed the way I think about my work and the work of others, he has and remains to be a major influence in my life and I hope we can continue working together in the future. I am grateful for Mario Falchi's views and opinions on the direction and impact of our work, which, at times admittedly, has been over-analysed.

I am grateful to current and past students and staff at the Department of Twin Research, without which the work presented in this thesis would not have been possible. All twin data presented here originates from historical twin visits and were collected and processed by the clinical research and lab teams led by Gail Clement and Gabriela Surdulescu respectively. Blood samples from the visits were analysed by the haematology laboratory at St Thomas' Hospital. Genome-wide SNP data was prepared for generalised use by Massimo Mangino, Nicole Soranzo, Guangju Zhai, and Pirro Hysi. I would personally like to thank Andrew Anastasiou for some very interesting conversations and for enduring my vocal grievances with R packages and all the joys their dependencies cause. Irina Gillham-Nasenya for her almost super-human ability for all things to do with the phenotype database. Raj Gill, Kathrine Hilario, and Karolina Zlobecka for their administrative support during our mitochondrial postal study.

I am also indebted to Nikolas Maniatis, Winston Lau, and Heather Elding for their generous hospitality in allowing me into their research group at University College London and the opportunity to obtain early access to their genomic data analyses.

Finally, I would thank my family for their continuing support and encouragement.

This PhD was funded by the Medical Research Council (50%) and King’s College London (50%).

Abstract

This thesis explores three broad areas of interest in the pathophysiology of type 2 diabetes (T2D). The first part of the thesis examines the relative contributions of body fat measurements on T2D and related morbidities in a large cohort of twins. A proxy measure of visceral fat was constructed from anthropometric and dual-energy X-ray absorptiometry, its heritability was estimated at 58% using the classical twin model and its influence on morbidity was compared to total abdominal fat and the body mass index. The findings from this work show that intra-abdominal adiposity confers the greatest independent risk on morbidity and appears to almost entirely mediate the observed association between morbidity and all the other measures of adiposity investigated.

The second part of this thesis is a candidate gene study of the *PARL*/*ABCC5* gene region motivated by prior evidence suggesting a role for *PARL* in T2D susceptibility. Using a single marker test of association, SNPs in and around the *PARL* gene showed no evidence of association with T2D. However, analysis based upon SNPs in the entire gene region (184,743-185,548Kb, build 36) using a multi-marker test of association, provided strong evidence that the neighbouring gene (*ABCC5*) is associated with T2D in both European and African American samples. In addition, *ABCC5* expression in subcutaneous adipose tissue was strongly associated with fasting insulin and glucose serum levels, visceral fat accumulation, and T2D with evidence that the disease susceptibility variant(s) is a regulatory element (an expression quantitative trait locus) located at intron 26 in *ABCC5*.

The third component of this thesis is a comprehensive investigation into the potential role of nuclear-encoded mitochondrial (NEM) genes in the aetiology of T2D. A pathway analysis approach is used to test for enrichment of T2D association signals across the genome in defined NEM gene sets. From this analysis, the biological pathways of glycolysis, the tricarboxylic acid cycle, and mitochondrial translation all show evidence of pathway enrichment. These findings demonstrate for the first time, potential associations between these pathways and T2D susceptibility in European and African American samples.

List of Tables

List of Figures

Abbreviations

2-h PG	2 hour plasma glucose
AA	African American
ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5
ADP	Adenosine biphosphate
ALK	Alkaline phosphatase
ALT	Alanine transaminase
ATP	Adenosine triphosphate
BC	Body cavity
BIL	Bilirubin
BMI	Body mass index
bp	Base pairs
BP	Blood pressure
CDCV	Common disease common variant
CDRV	Common disease rare variant
cIMT	Carotid intima-media thickness
CSA	Cross sectional area
CT	Computed tomography
CVD	Cardiovascular disease
DAVID	Database for Annotation, Visualization and Integrated Discovery
dbGAP	Database of Genotypes and Phenotypes
DGI	Diabetes Genetics Initiative
DNA	Deoxyribonucleic acid
DTR	Department of Twin Research
DXA	Dual-energy X-ray absorptiometry
DZ	Dizygotic
EEA	Equal environment assumption
eQTL	Expression quantitative trait locus
ESRD	End stage renal disease
FFA	Free fatty acids
FHS	Framingham Heart Study
FPG	Fasting plasma glucose
GGT	Gamma-glutamyl transpeptidase
GO	Gene Ontology
GSEA	Gene set enrichment analysis
GWA	Genome-wide association
HATS	Healthy Ageing Twin Study
HbA1c	Glycosylated haemoglobin
HOMA	Homeostasis Model Assessment
HOMA-%B	Homeostasis Model Assessment beta cell function
HOMA-%S	Homeostasis Model Assessment insulin sensitivity
HOMA-IR	Homeostasis Model Assessment insulin resistance
HR	Hazard ratio

HT	Hypertension
IGR	Insulin glucose ratio
IGT	Impaired glucose tolerance
IID	Identical, independent distributions
Kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCL	Lymphoblastoid Cell Line
LDU	Linkage disequilibrium units
LFT	Liver function test
LRT	Likelihood ratio test
MAF	Minor allele frequency
MAGENTA	Meta-Analysis Gene-set Enrichment of variaNT Associations
MES	Maximum enrichment score
MIDD	Maternally Inherited Diabetes and Deafness
MIPS	Munich Information Center for Protein Sequences
MODY	Maturity Onset Diabetes of the Young
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
MSigDB	Molecular Signatures Database
mtDNA	Mitochondrial deoxyribonucleic acid
MuTHER	Multiple Tissue Human Expression Resource
MZ	Monozygotic
NaDIA	National Diabetes Audit
NCBI	National Center for Biotechnology Information
NEM	Nuclear-encoded mitochondrial
NHGRI	National Human Genome Research Institute
NIDDK	National Institute of Diabetes and Digestive and Kidney diseases
OGTT	Oral glucose tolerance test
OLS	Ordinary least squares
OR	Odds ratio
OXPHOS	Oxidative phosphorylation
PARL	Presenilin associated, rhomboid-like
QC	Quality control
QOF	Quality and Outcomes Framework
QUICKI	Quantitative insulin sensitivity check index
ROC	Receiver operating curve
rRNA	Ribosomal ribonucleic acid
SFW	Subcutaneous fat width
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TCA	Tricarboxylic acid
TED	Transverse external diameter

TID	Transverse internal diameter
tRNA	Transfer ribonucleic acid
UKBS	United Kingdom Blood Service
VAT	Visceral adipose tissue
VIF	Variance inflation factor
WC	Waist circumference
WHO	World Health Organisation
WTCCC	Wellcome Trust Case Control Consortium

Publications

Direk K, Cecelja M, Astle W., Chowienczyk P, Spector T, Falchi M, Andrew T (2013) The relationship between DXA-based and anthropometric measures of visceral fat and morbidity in women. *BMC Cardiovascular Disorders* **13**:25.

Direk K, Lau W, Small KS, Maniatis N, Andrew T (2013) *ABCC5* Transporter is a Novel Type 2 Diabetes Susceptibility Gene for European and African American Populations. *Submitted*.

Chapter 1: Introduction

This chapter provides the necessary background information to the research topics discussed in this thesis. I begin by defining type 2 diabetes (T2D) and describing how this disease is diagnosed and its prevalence both worldwide and in the United Kingdom. In the second half of the introduction, I review the supporting evidence for a mitochondrial contribution to T2D susceptibility and end with an outline of the research aims of this thesis.

Type 2 Diabetes

It is a condition of chronic hyperglycaemia as a result of insulin resistance and pancreatic beta cell dysfunction - insulin resistance describes the biological lack of sensitivity to the peptide hormone insulin secreted by the pancreatic beta cells. When beta cells are not able to secrete sufficient quantities of insulin in order to reduce glucose level, this is termed beta cell dysfunction. Either extreme insulin resistance or beta cell dysfunction is sufficient to cause diabetes – the depletion of pancreatic beta cells in type 1 diabetes (T1D) can be considered an extreme from of pancreatic beta cell dysfunction. Moreover, rare and severe forms of insulin resistance are known (Mantzoros *et al.*, 1998, Semple *et al.*, 2011). With the development of T2D, the increasing lack of tissue sensitivity to insulin and compensatory increase in insulin secretion is not adequate to decrease the blood glucose levels, which if left unchecked, is toxic and ultimately results in irreversible nerve damage and organ failure.

Himsworth (1936) first recognised diabetic individuals could be classified into insulin sensitive and resistant phenotypes. The first report by the World Health Organisation (WHO) on diabetes in 1964 initially classified diabetes by age of onset; this was later changed in their 1980 report by distinguishing the two major types of diabetes by insulin sensitivity. As more was learnt about diabetes, new diagnostic methods were introduced along with re-evaluation of the diagnostic criteria, the latter being reclassified on many occasions, which has an impact on historical epidemiological comparisons of prevalence and incidence. It was thought that insulin resistance was the primary pathogenic trigger in T2D, requiring tissues to demand ever-increasing amounts of insulin. It is now recognised that predisposition to beta cell dysfunction is likely to be a pre-existing condition within individuals that go on to develop T2D - insulin secretion is ostensibly adequate in the context of high insulin sensitivity and the inability for compensatory insulin secretion is hidden at this stage. Only as peripheral sensitivity to insulin decreases, does this beta cell defect become apparent.

This current view has been is supported by a recent study from Tabak *et al.* (2009), in which they investigated the relationship between insulin resistance and pancreatic beta cell function prior to T2D onset using longitudinal data collected for approximately a decade on more than 6,500 British civil servants. The study confirmed the known inverse relationship between age and insulin sensitivity – as age increases, insulin sensitivity decreases in both non-diabetics and those who developed diabetes during the course of the study (incident diabetes). In the non-diabetics, beta cell function remained essentially constant with increasing age. However, as shown in Figure 1.1, those individuals that go on to develop T2D have higher beta cell function than the non-diabetic group for almost the entire length of the study. This is combined with poorer insulin sensitivity for the same period and a brief compensatory increase followed by an abrupt decline in beta cell function shortly (~3 years) prior to T2D diagnosis. This study supports the concept of a primary pancreatic beta cell defect, which when confronted with decreasing insulin sensitivity, temporarily increases insulin output followed by decompensation of beta cells (Weir and Bonner-Weir, 2004).

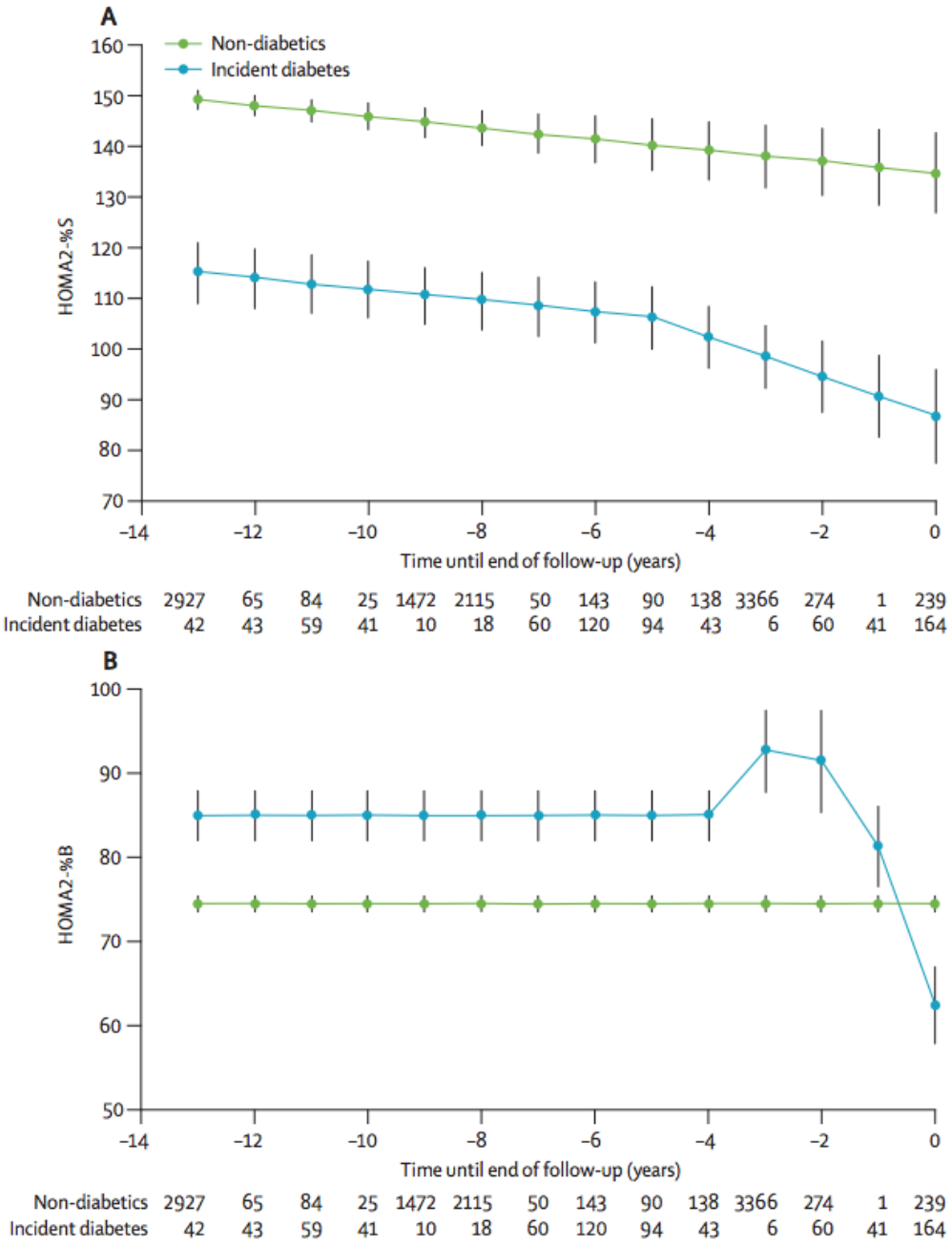


Figure 1.1. Insulin sensitivity (A) and beta cell function (B) time course to T2D diagnosis or end of follow up for non-diabetics (Tabak *et al.*, 2009).
Graphs A and B show a 13-year prospective time course comparison of beta cell function and insulin sensitivity in those that become diabetic (blue line) and those that do not (green line). Incident cases of diabetic and non-diabetic at each time point are given below graphs A and B.

Insulin and Glucose Action

In normal physiology, increased blood glucose causes an array of metabolic responses, one of which is an increase in the rate of insulin secretion from pancreatic beta cells. Notable targets of insulin action include the liver, muscle, and adipose tissue, where insulin stimulates glucose uptake (Leney and Tavare, 2009), while suppressing endogenous glucose production in the hepatocytes (Postic *et al.*, 2004). The cumulative effect of insulin in the presence of hyperglycaemia therefore acts to lower the concentration of circulating glucose, returning to a state of normoglycaemia. Whilst beta cells are able to increase insulin secretion in response to glucose, they eventually reach a plateau in which insulin output is not able to restore normoglycaemia.

Insulin secretion by the pancreatic beta cell is largely driven by the ATP to ADP ratio within the cell. Known as the triggering pathway – an increase in the ATP:ADP ratio causes ATP-sensitive potassium channels to close, depolarising the cell membrane, which opens the calcium channels and the influx of calcium into the cell triggers the exocytosis of insulin-containing vesicles. An additional, potassium-channel-independent amplifying pathway has been described (Henquin, 2011), in which glucose-mediated insulin secretion occurs when ATP-sensitive potassium channels are prevented from either opening or closing, through the use of diazoxide and sulfonylurea. Other studies (Nenquin *et al.*, 2004, Ravier *et al.*, 2009) have also shown insulin secretion occurs in beta cells lacking components of the ATP-sensitive potassium channel.

As the major producer of ATP, mitochondria can be considered central to the triggering pathway of insulin secretion. Anaplerotic products from mitochondrial metabolic activities, such as glutamate and malonyl-CoA are also implicated in the amplifying pathway (Maechler *et al.*, 2006). Defects in mitochondrial processes that are involved in ATP production leading to failure to attain a sufficient ATP:ADP ratio have the potential to contribute to beta cell dysfunction in T2D. One of the most effective oral agents for T2D (Halimi, 2006) is metformin, (1,1-dimethylbiguanide). Pertinently, its primary mechanism of action is inhibition of the mitochondrial protein complex 1 (Owen *et al.*, 2000), which as a result, reduces hepatic gluconeogenesis (Foretz *et al.*, 2010) leading to a reduction in circulating glucose.

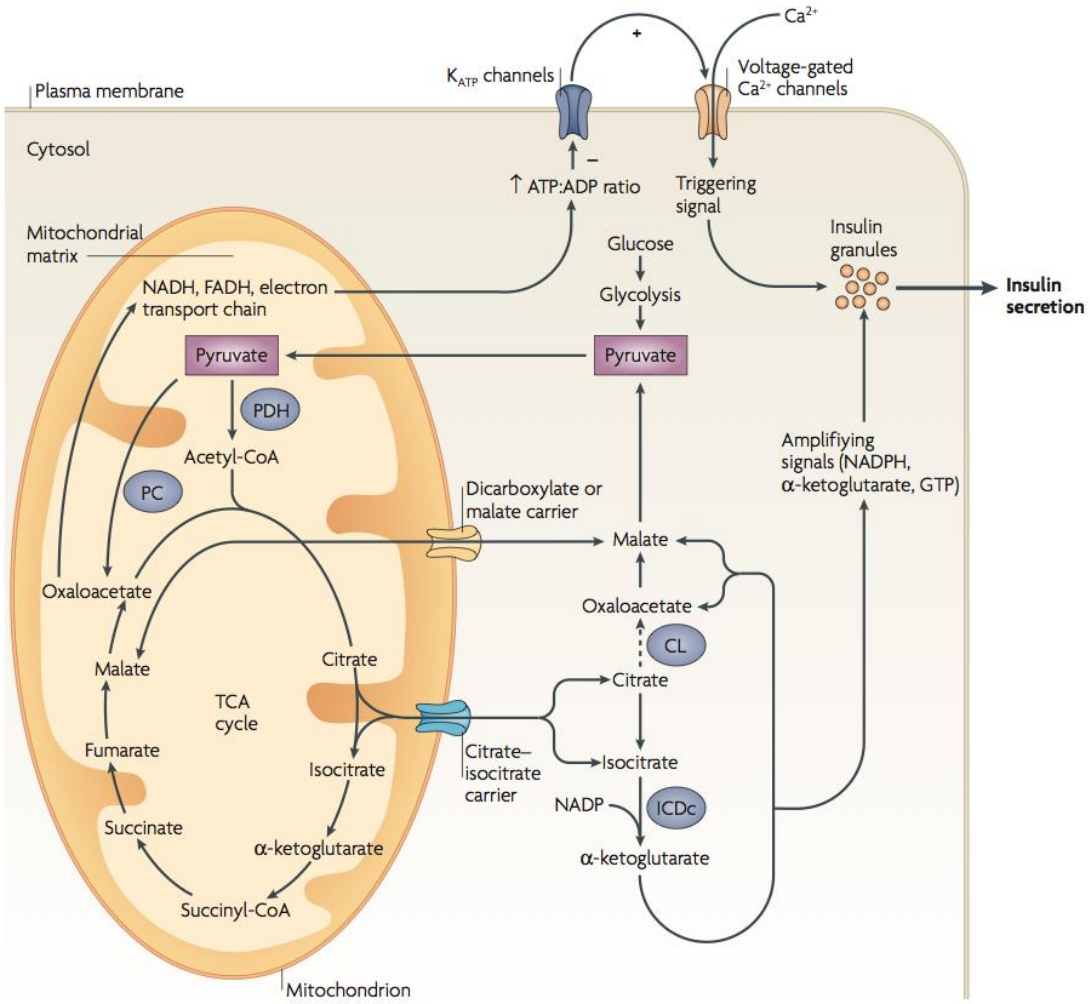


Figure 1.2. Molecular pathways controlling glucose-stimulated insulin secretion in pancreatic beta cells (Muoio and Newgard, 2008).

Products of glucose metabolism, in addition to the adenosine triphosphate (ATP) to adenosine biphosphate (ADP) ratio, trigger mitochondrially-derived signals and promote the secretion of insulin from pancreatic beta cells.

Measures of Insulin Resistance and Beta Cell Function

An important measure in describing insulin sensitivity is the ability to model the physiological relationship between glucose and insulin, assessing the combined contributions of both beta cell function and insulin resistance. For this purpose, the homeostasis model assessment (HOMA) was described by Matthews *et al.* (1985), the model uses fasting plasma insulin and glucose concentrations to estimate the degree of beta cell function and insulin resistance as shown by two formulae:

$$\text{HOMA1-\%B} = (20 \times \text{fasting plasma insulin}) / (\text{fasting plasma glucose} - 3.5)$$

$$\text{HOMA1-IR} = (\text{fasting plasma insulin} \times \text{fasting plasma glucose}) / 22.5$$

Where, HOMA1-%B and HOMA1-IR refer to beta cell function and insulin resistance respectively; a third measure, HOMA1-%S is an estimate of insulin sensitivity and equal to 100 / HOMA1-IR. The units of measure used for this model are milliunits per litre (mU/l) for fasting plasma insulin and millimoles per litre (mmol/l) for glucose concentration. For insulin, 1 mU/l is equal to 6.945 pmol/l.

The original HOMA was later updated by Levy *et al.* (1998) to incorporate a more accurate representation of the relationship over a range of insulin and glucose values. The updated version (HOMA2) differs from the original HOMA in two key areas; the first accounts for hyperglycaemic suppression of hepatic glucose production, and the second allows a variable insulin secretion response to glucose concentrations (Wallace *et al.*, 2004). Thus, HOMA2 is the recommended model for use (Levy *et al.*, 1998). For both HOMA versions, beta cell function and insulin resistance cannot be reported in isolation as it may lead to erroneous interpretation (Wallace *et al.*, 2004).

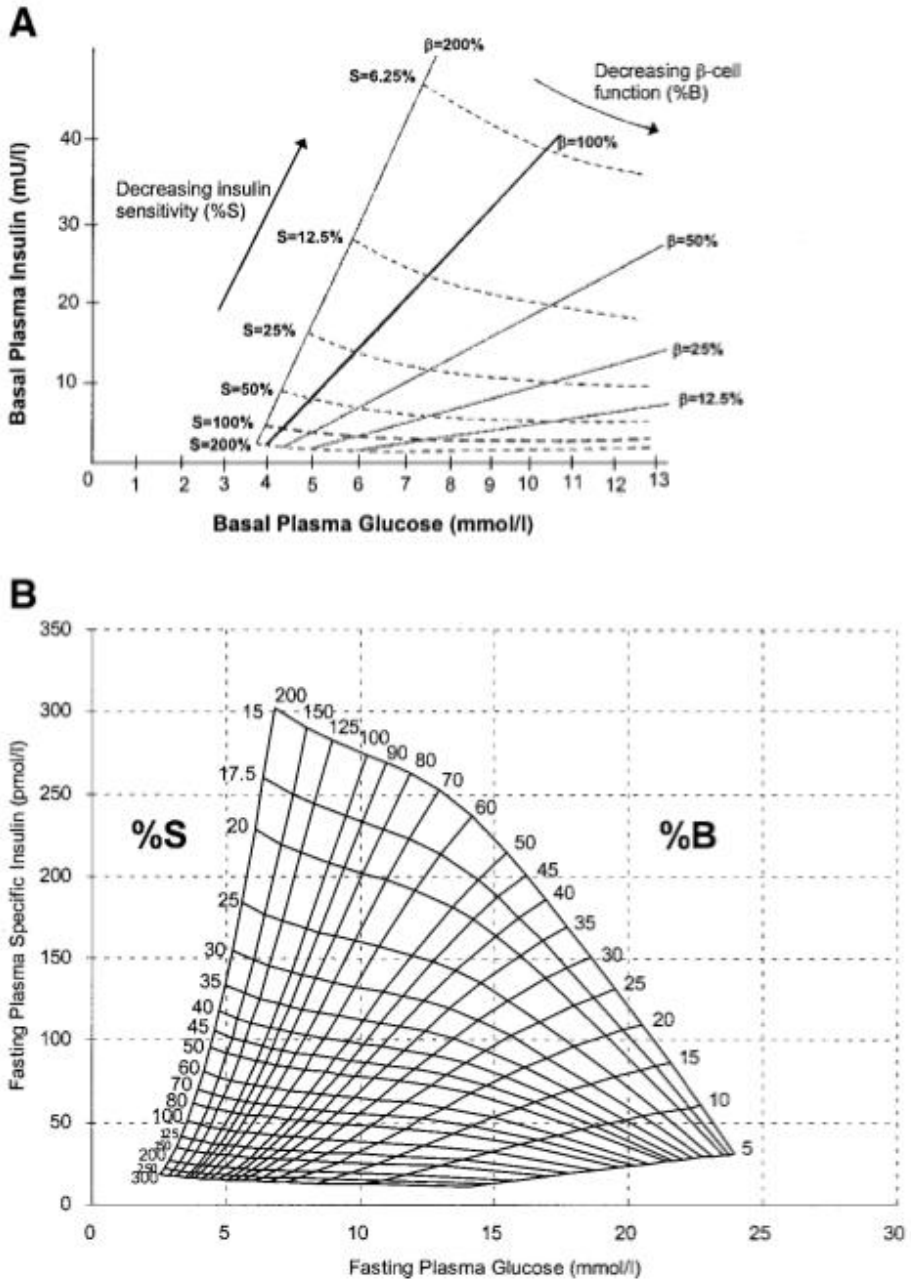


Figure 1. 3. Original (A; HOMA) and the updated (B; HOMA2) versions of the homeostasis model assessment (Wallace *et al.*, 2004).

Graphs illustrate the relationship between beta cell function (%B) and insulin sensitivity (%S) plotted over a range of fasting plasma glucose (x axis) and insulin (y axis) values. The original model (A) is linear, whilst the updated model (B) is curvilinear representing the variable secretion of insulin in response to glucose concentrations.

Both HOMA1 and HOMA2 estimates for beta cell function and insulin resistance have shown to be correlated with the same measures obtained through the euglycaemic and hyperglycaemic clamp tests. This is noteworthy because the HOMA estimates beta cell function and insulin resistance under normal physiological conditions for insulin and glucose, whereas, the euglycaemic and hyperglycaemic clamp tests specifically alter these parameters way and measure the response. To appreciate the difference, it is important to briefly outline the principles behind the clamp-based methods.

The insulin and glucose clamp techniques was first described over 30 years ago by DeFronzo *et al.* (1979), the hyperglycaemic clamp involves raising the plasma glucose concentration intravenously to reach a hyperglycaemic equilibrium of 125 mg/dl (6.9 mmol/l) above basal glucose level, once equilibrium is obtained, it maintained for 2 hours through infusion of glucose "maintenance doses". Under such conditions, glucose infusion is a measure of insulin secretion. In the euglycaemic clamp, the plasma insulin concentration is raised to 100 μU/ml (694.5 pmol/l) above basal insulin level and maintained at this level by the infusion of a fixed dose of insulin. Shortly after, the plasma glucose concentration is held at a basal level by variable glucose infusion. In this state, the rate of glucose infusion is an indicator of insulin clearance, thus sensitivity.

Wallace *et al.* (2004) have argued that HOMA should be used where possible instead of other indices such as the fasting insulin to glucose ratio (IGR) or the quantitative insulin sensitivity check index (QUICKI; 1/[log (fasting plasma insulin x fasting plasma glucose)]). QUICKI is effectively the log of HOMA1-IR. Discriminatory power of these indices to identify insulin resistance is usually determined from receiver operator curve (ROC) analysis; individuals that are truly insulin resistant are classified by direct tests such as the hyperinsulinemic-euglycaemic clamp. For a thorough review of current methods to assess insulin resistance and their indices please see Muniyappa *et al.* (2008).

In the “Uses Abuses of HOMA Modeling” paper, Wallace *et al.* (2004) note an important caveat about how estimates from HOMA should be interpreted in relation to the clamp-based methods, in that the two techniques “give information about different aspects of beta cell function or insulin resistance”. For epidemiological purposes, this point is pertinent as invasive clamp methods are unfeasible for large studies and so insulin-glucose indices are used, justified by correlation with clamp methods that operate outside the normal

physiological range of insulin and glucose. In fact, studies incorporating these clamp-based methods cannot be compared if infusion rates differ (Natali *et al.*, 2000). Moreover, no one metric combining insulin and glucose that outperforms another in terms of correlation with euglycaemic clamp-measured insulin resistance (Ruige *et al.*, 2006, Lorenzo *et al.*, 2010). These results suggest that surrogate measures of insulin resistance provide the same or very similar information on insulin resistance; therefore, with careful interpretation, the choice between them might not be too important for research purposes.

In addition, Samaras *et al.* (2006) have rightly pointed out that attention should instead be focused on how diabetes risk factors influence insulin resistance (McLaughlin *et al.*, 2007). Cut-off values to identify insulin resistance, where they exist, are arbitrary, population and assay specific; furthermore, insulin resistance is a continuous scale – every individual has some level of insulin resistance. It is unnecessary to dichotomise a measure of insulin resistance where there is no evidence to suggest that individuals that sit immediately adjacent to the cut-off point differ substantially in future risk of morbidity. However, if the entire distribution of insulin resistance is considered, meaningful comparisons can be made between the tails of the distribution in the context of risk factors.

An exception to this argument is made for disease diagnosis based on a quantitative trait such as glucose concentration for T2D, body mass index (BMI) for obesity, or blood pressure for hypertension. Cut-off values for these disorders are somewhat misleading in that they should identify individuals with a substantially increased risk of future mortality and complications, one such example is the steep increase in retinopathy prevalence for individuals above the diagnostic criteria for T2D (Figure 1.4). There is no dispute that individuals with extremely high values will experience poor future health, but for those individuals that flank the threshold, the difference in future morbidity is less clear, reflecting the attempt of extrapolating a single continuous measure to future organ or system-level dysfunction. Diagnosis for T2D differs from that of obesity and hypertension because the measurement, i.e. glucose level, is of that of the perpetrator; a high glucose concentration has a direct detrimental effect on cells.

Diagnostic Test Criteria for T2D

Many people with T2D are often unaware due to its asymptomatic nature in the early phase of the disorder; symptoms such as thirst, frequent urination, and weight loss can go unnoticed. Diagnosis of T2D is based on a variety of tests that measure glucose concentration. I will briefly outline the most common measures and comment on their qualities, as these measures are available for some of our study sample.

One method to identify hyperglycaemic individuals is to measure the amount of glucose in blood plasma after eight or more hours of fasting. This was the most popular method of diagnosing T2D due to its simplicity and low cost relative to other methods. The fasting plasma glucose (FPG) test measures the product of two opposing actions – endogenous glucose production, which replenishes glucose level during a period of fasting, and the response by insulin to lower elevations in glucose level. Despite its popularity, the FPG test suffers from major limitations: the individual must fast for the preceding eight hours before the test, two tests are required to confirm a positive diagnosis, and FPG is influenced by many behavioural and biological factors resulting in within- and -between day variability. Lastly, it has been noted (Schrot *et al.*, 2007) that glucose measurement in blood plasma is usually higher than the same measure in whole blood.

Another method to measure blood glucose level is by glycosylated haemoglobin (HbA1c). The current guidelines on HbA1c as a diagnostic test are presented in the 2011 WHO diabetes report (http://www.who.int/diabetes/publications/report-hba1c_2011.pdf). The report made the recommendation that a HbA1c level greater than 6.5% (47.5 mmol/mol) can be used to diagnose T2D and that values less than this do not exclude diabetes diagnosis from glucose tests. HbA1c is the preferred method of T2D diagnosis as it captures the average plasma glucose over the preceding 3 months, does not require the individual to be in the fasting state and has less day-to-day variability than FPG. The HbA1c and FPG diagnostic criteria for T2D are comparable for prevalence of microvascular complication, such as moderate non-proliferative retinopathy (Figure 1.4).

Measure	Diagnostic criteria
Glycosylated haemoglobin (HbA1c)	>6.5%
Fasting plasma glucose (FPG)	>7.0 mmol/l
2-h plasma glucose	>11.1 mmol/l
Random plasma glucose	>11.1 mmol/l

Table 1.1. The American Diabetes Association diagnostic criteria for T2D (American Diabetes, 2011)

This table lists four diagnostic criteria for T2D; repeated tests that exceed any of these criteria confirm T2D diagnosis. Fasting plasma glucose is measured after the individual has fasted for at least eight hours; the non-fasted alternative measures random plasma glucose and has the same criterion as measuring glucose 2 hours (2-h) after an oral glucose challenge.

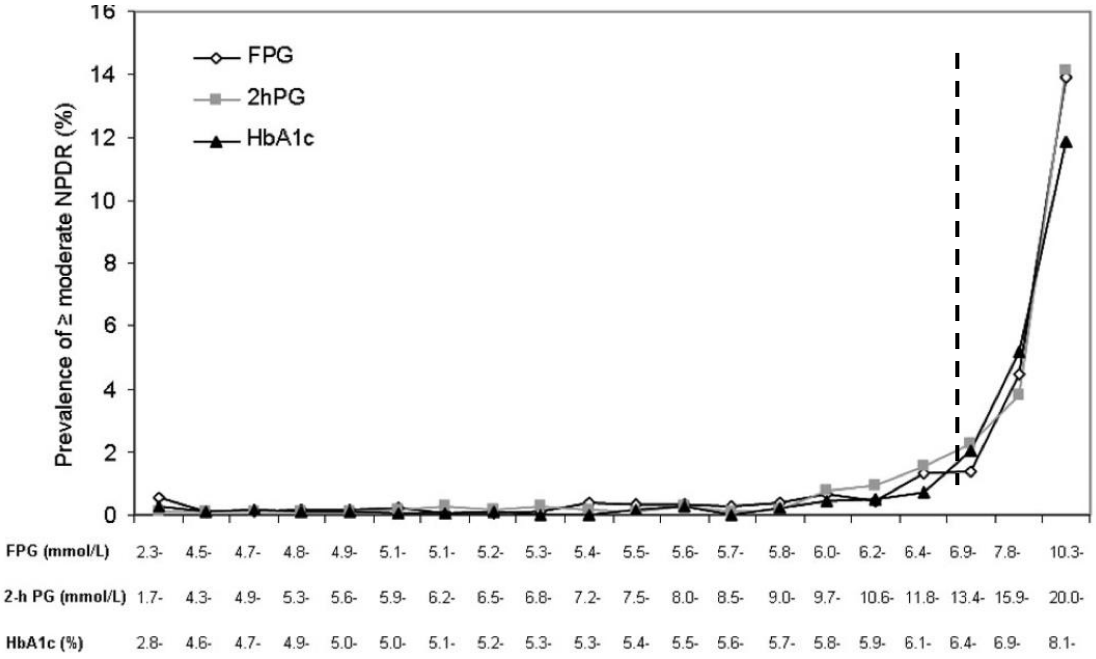


Figure 1.4. Prevalence of moderate non-proliferative retinopathy (NPDR) by vigintiles of fasting plasma glucose (FPG), 2-hour plasma glucose (2-h PG), and glycosylated haemoglobin (HbA1c).

The prevalence of NPDR increases rapidly above glycaemic diagnostic thresholds (indicated by the vertical hashed line) for T2D. Adapted from the WHO Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus (2011).

Prevalence and Incidence

The prevalence of T2D is estimated by the diagnostic measure of glycaemia used for diagnosis. Diagnostic criteria (Table 1.1) have fluctuated throughout history reflecting epidemiological data on micro/macrovascular complications for hyperglycaemic individuals (Figure 1.4).

Worldwide

It is estimated that 336 million people are affected by T2D (Ashcroft and Rorsman, 2012, Danaei *et al.*, 2011), the most common form of diabetes, accounting for 90-95% of all cases. Other types of diabetes included T1D, which is predominantly caused by immune-mediated destruction of beta cells, and gestational diabetes, which occurs during pregnancy.

England

Diabetes statistics in England and Wales is publically available by The National Diabetes Audit (NDA) and Diabetes UK for the National Health Service (NHS). Table 1.2. shows year-on-year increases in the prevalence of diabetes in England from 2004 to 2011. The most recent report by the NDA is the 2010-2011 report, which for England, combined diabetes patient data from 152 Primary Care Trusts (2.15 million individuals of an estimated 2.46 million diagnosed with diabetes). The NDA report from the previous year (2009-2010) provided a gender and ethnicity breakdown of T2D prevalence; diagnosis of T2D in males was higher than females (4.35% vs. 3.41%) and individuals that describe themselves as white made up 77.7% of the 1.9 million T2D cases in this report. The most recent report (2013) by Diabetes UK (http://www.diabetes.org.uk) estimates the number of individuals diagnosed with T2D at 2.56 million and a diabetes prevalence of 5.5% in England, and if the ratio of T2D to T1D is 1:10 as stated in the report; the 2011 prevalence of T2D in England estimated to be 4.95%.

Year	Sample Size (Millions)	Total Diabetes Diagnosed (Millions)	Prevalence (T2D Prevalence)	Source
2004	-	1.48	2.95 (2.66)	Diabetes UK 2004
2005	-	2.10	3.50 (3.15)	Diabetes UK 2005
2006	-	2.20	3.73 (3.35)	QOF/Diabetes UK
2007-2008	1.41	2.09	3.91 (3.52)	Extrapolated from NDA 2008-2009
2008-2009	1.70	2.20	4.13 (3.65)	NDA 2008-2009
2009-2010	1.9	2.34	4.35 (3.87)	NDA 2009-2010
2010-2011	2.15	2.46	4.57 (4.07)	NDA 2010-2011

Table 1.2. Diabetes prevalence in England between 2004-2011.
Summary statistics from the National Diabetes Audit for adults. Total diagnosed from Quality and Outcomes Framework (QOF; http://www.qof.ic.nhs.uk). Where there is discrepancy between QOF/Diabetes UK and the National Diabetes Audit (NDA), NDA values are presented as these are better documented. Note that pre-2007, prevalence estimates were not given and so calculated using the Office for National Statistics (ONS) data: http://www.ons.gov.uk/ons/publications/all-releases.html?definition=tcm:77-22371.

Mitochondrial Function and T2D

In order to investigate mitochondrial function in relation to insulin resistance and T2D, it is important to look at the appropriate tissue, i.e. tissues that are major sites of insulin-stimulated glucose production, uptake, and/or disposal. The major tissues in this category are the pancreas, the liver, skeletal muscle, and adipose tissue. Dysfunctional mitochondria or low mitochondrial content in the cells of these tissues may have a direct consequences on insulin secretion or glucose management that depend on mitochondrial processes. I will outline evidence for and against this hypothesis in the following sections.

Adipocytes are the main storage site for free fatty acids (FFA) in the human body. Subcutaneous adipose tissue is responsible for much of the FFA in circulation (Karpe *et al.*, 2011) while visceral adipose tissue seems to be the prominent source of hepatic FFA delivery (Nielsen *et al.*, 2004). Any comparisons in mitochondrial content between subcutaneous and visceral adipose tissues or subcutaneous/visceral adipose tissue between normoglycaemic and T2D individuals, is complicated by the fact that adipose tissue contains adipocytes of different sizes and thus a given tissue mass may contain an unequal number of adipocytes.

In rats, it is known that accumulation of visceral adiposity follows stages of hypertrophy, an increase in adipose cell count, and reduction in cell size (Strissel *et al.*, 2007, Leney and Tavare, 2009). There is little reason why this time course should not also be true for humans, but without estimates of the proportions of adipocyte size, it is difficult to reach consensus over mitochondrial content and activity between studies.

Laye *et al.* (2009) tracked changes in mitochondrial content in visceral adipose tissue from normoglycaemic to T2D showing mitochondrial content decreases in obese T2D rats compared to non-obese normoglycaemic rats. This finding is consistent with earlier work (Choo *et al.*, 2006), which showed in both T2D (leptin receptor deficient, db/db) and obesity (leptin deficient, ob/ob) mouse models have fewer and smaller mitochondria in adipocytes than their heterozygous or wild-type counterparts.

Increased circulating FFA have been shown to increase the mitochondrial content of muscle cells along with the rate of fatty acid oxidation and development of insulin resistance (Hancock *et al.*, 2008, Koves *et al.*, 2008). Importantly these studies highlight the adaptive

response by mitochondria to oxidise fatty acids, which is known to impair insulin signalling transduction pathways (Dresner *et al.*, 1999, Virkamaki *et al.*, 2001, Yu *et al.*, 2002). However, this does not contradict the observations that oxidative phosphorylation activity in muscle cells from T2D individuals is reduced compared to controls (Phielix *et al.*, 2008), mitochondria are smaller in muscles obtained from T2D individuals (Kelley *et al.*, 2002), and insulin resistance correlates with mitochondrial content within muscle fibres (Chomentowski *et al.*, 2011).

Genetics of T2D

Genetic Architecture of Common Disease

T2D has a multifactorial aetiology; both genetic and environmental factors contribute to the trait. A critical factor in unravelling the genetic component and successfully identifying susceptibility variants is the mode of inheritance or the genetic architecture of the trait; that is the number of gene loci involved, their allele frequency and penetrance. The hope for gene mapping studies of complex traits is that the mode of inheritance includes at least some genetic loci (oligogenic) of moderate effect which are detectable, in addition to the many loci (polygenic) with small effect.

Genome-Wide Association Studies

The common disease common variant (CDCV) hypothesis posits that disease-predisposing alleles with relatively high frequency in the population underlie genetic susceptibility to common diseases (Reich and Lander, 2001). The same idea can be applied to any quantitative trait, such as height or BMI. The CDCV assumption underpins the study design of genome-wide association (GWA) studies; common alleles that confer susceptibility to the trait are expected to be present at a higher frequency in the affected/disease group compared to the unaffected/control group.

GWA studies as a method of genetic analysis have empirically confirmed that most complex traits appear to have a polygenic mode of inheritance (Yang *et al.*, 2010, Clayton, 2009). Using T2D and related glycaemic traits as an example, over 60 loci (Morris *et al.*, 2012) have been identified as conferring susceptibility to these traits. Lack of statistical power due to small effect sizes and lack of linkage disequilibrium (Chapter 2: Linkage Disequilibrium) between the causal and genotyped single nucleotide polymorphism (SNP), combined with stringent significance threshold incurred for testing hundreds of thousands to millions of SNPs may hide many more loci (Spencer *et al.*, 2009). Conventional GWA studies use a significance threshold of 10^{-8}, which is based on an estimate for the effective number of tests performed in a GWA scan (Tanaka, 2005, Dudbridge and Gusnanto, 2008). This figure is similar to a standard Bonferroni correction, in which the desired nominal significance for a single test is divided by the total number of tests performed. For example, if the desired alpha value for a single test is 0.05, the significance threshold for testing 500,000 SNPs is a p-value of $1x10^{-7}$ (0.05/500,000).

Eight years after the first GWA study (Klein *et al.*, 2005) over 2000 SNPs have been associated with complex traits (Visscher *et al.*, 2012), yet for any particular trait, the implicated genetic variants explain a smaller fraction of the genetic component of the trait, as estimated from twin or family studies (Table 1.3). This situation has been termed as "missing [narrow-sense] heritability" and explanations have been extensively discussed (Manolio *et al.*, 2009, Eichler *et al.*, 2010). As of 2009, the 18 loci identified for T2D explained just 6% of its narrow-sense heritability, approximately the same figure was found for height with more than double the number of loci, in comparison, some 50% of the narrow-sense heritability for age-related macular degeneration is explained by just 5 loci (Manolio *et al.*, 2009). Potential sources of "missing heritability" include rare variants with large effects (CDRV hypothesis) and/or the effect of structural variants; however, it is not due epigenetics (Slatkin, 2009), nor gene-gene interactions or gene-environment interactions as these do not contribute to additive genetic variance (Chapter 2: Variance).

Peter Visscher and colleagues provide a compelling argument in this discussion. In the first of two studies by the group, Yang *et al.* (2010) explore an alternative method to test for SNP-trait associations, whereby instead of testing each SNP individually, they fit a linear model to estimate how much phenotypic variance is explain by considering all genomic SNPs simultaneously using REsticted Maximum Likelihood (REML) regression models. Using this approach, genomic SNPs collectively accounted for 45% of the phenotypic variance; this estimate of the phenotypic variance explained by empirical SNP data falls short of the frequently reported narrow-sense heritability of 80% for height.

One reason for this discrepancy is that LD between genotyped SNPs and causal variants is incomplete; for example, if LD between genotyped common and genotyped rarer SNPs is comparable to LD between (non-genotyped) causal variants and genotyped rarer SNPs, then the empirical SNP-based narrow-sense heritability increases to 54%. However, if the minor allele frequency (MAF) of the causal variants are much lower than the genotyped SNPs (e.g. an allele spectrum of intermediate and rare allele frequencies), LD would be substantially reduced between the two and the proportion of variance explained by genotyped SNPs would rise to 80%, thereby accounting for the entire narrow-sense heritability of height. Hence the "missing heritability" for height can be completely accounted for by a spectrum of common and rarer alleles, which may or may not include rare (MAF < 0.01) causal alleles.

The second study (Yang *et al.*, 2011) related to the subject of "missing heritability" in relation to four traits, including height and BMI to examine how the phenotypic variance is captured by common genotyped SNPs distributed throughout the genome. Across these traits, the amount of phenotypic variance explained by genotyped SNPs is (1) much higher than previously thought and (2) positively associated with the number of genes and physical length of the chromosome. This latter point is relevant because, as the authors comment, if genetic variation is spatially correlated with genic regions, then these regions are prime candidates for harbouring causal variants, and depending upon the trait, provides justification for using exome capture as opposed to whole-genome sequencing in the search for causal variants.

Figure 1.5 conceptually illustrates the wide range of allelic spectra underpinning multi-factorial and monogenic disease; much of the attention to genetic variants in this introduction and this thesis is towards common variants, which are present at frequencies greater than 5% in the population. Low frequency (>0.5% and <5%) and rare (<0.5%) genetic variants are now being explored in the hope that they will reveal more susceptibility loci for complex traits and with larger effects than have been witnessed for common variants. The 1000 genome project (http://www.1000genomes.org) is currently underway with the aim of cataloguing genetic variants down to a MAF of 1%; to date it has identified 22 million novel SNPs, bringing the total number of catalogued SNPs in the human genome to 38 million (Buchanan *et al.*, 2012).

It is unknown how many susceptibility loci remain to be found and to what extent do they influence the complex traits. To date, many of the loci identified for T2D have low effect size and penetrance – carrying any single risk allele or even all known risk alleles does not determine that an individual will develop T2D.

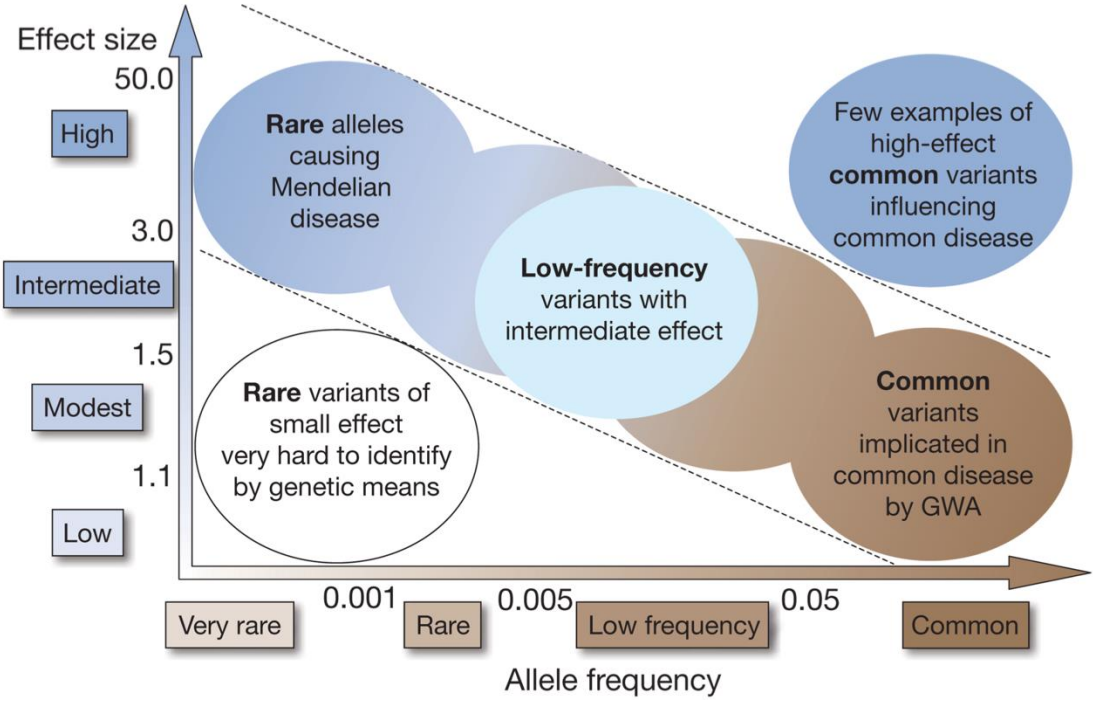


Figure 1.5. Spectrum of allele frequencies and effect sizes from (Manolio *et al.*, 2009).
In this graph various combinations exist for the frequency of a genetic variant (x-axis) and its effect size expressed as an odds ratio (OR, y-axis). Common variants with modest effect sizes are considered the low-hanging fruits in the search for genetic variants underlying complex traits. The stringent significance threshold at a genome-wide level, which can involve performing millions of tests, potentially conceals many more genetic variants with smaller effects.

Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits	h^2 All GWAS SNPs
Type 1 diabetes	0.90	0.60*	0.30
Type 2 diabetes	0.30 – 0.61	0.05 – 0.10	–
Obesity (BMI)	0.4 – 0.6	0.01 – 0.02	0.20
Crohn's disease	0.60 – 0.80	0.10	0.40
Ulcerative colitis	0.50	0.05	–
Multiple sclerosis	0.30 – 0.80	0.10	–
Ankylosing spondylitis	>0.90	0.20	–
Rheumatoid arthritis	0.60	–	–
Schizophrenia	0.70 – 0.80	0.01	0.30
Bipolar disorder	0.6 – 0.7	0.02	0.40
Breast cancer	0.30	0.08	
Von Willebrand factor	0.66 – 0.75	0.13	0.25
Height	0.80	0.10	0.50
Bone mineral density	0.60 – 0.80	0.05	–
QT interval	0.37– 0.60	0.07	0.20
HDL cholesterol	0.50	0.10	–
Platelet count	0.80	0.05 – 0.1	–

Table 1.3. Narrow-sense heritability and proportion of variance explained by SNPs (Visscher *et al.*, 2012).

The first column of the table shows the narrow-sense heritability (h^2) estimate from family studies, the second and third columns shows the proportion of trait variance explained by SNPs that reach genome-wide significance and all genotyped SNPs on the array respectively. * Includes pre-GWAS loci.

Monogenic Forms of Insulin Resistance

Severe monogenic forms of insulin resistance have been described, some affecting the whole body (George *et al.*, 2004, Savage *et al.*, 2002), whilst others are tissue specific. The original nomenclature used to describe the different forms of insulin resistance (Kahn *et al.*, 1976) discriminated two forms, type A and type B, where the primary defect in type A was insulin receptor binding, and the presence of anti-insulin receptor antibodies in type B. These subphenotypes of severe insulin resistance are now recognised to include all disorders of insulin signal transduction or defects in adipose tissue development or function (lipodystrophies), which may or may not have an immune-mediated component. A new classification system for diagnosis of severe insulin resistance has been proposed (Semple *et al.*, 2011), which delineates the aetiology between the two main forms as primary insulin-signalling defects or because of adipose tissue abnormalities.

Novel Biological Insights of Insulin Secretion/Resistance From GWA Studies

To date, approximately 60 (Morris *et al.*, 2012) T2D susceptibility loci have been identified. Table 1.4 demonstrates the impact of GWA studies on number of loci identified for T2D and related traits (NB. does not include the 10 additional loci identified by Morris *et al.* (2012) for T2D). Many of the reported genes for T2D encode proteins implicated in beta cell function, while disproportionally few genes are implicated in insulin resistance (Table 1.5). Potential reasons for the lack of identified insulin resistance genes have been discussed (Petrie *et al.*, 2011) and the main arguments are listed below:

1. Susceptibility variants in insulin resistance genes may have small effect sizes requiring very large samples.
2. Insulin secretion is limited to a single tissue with a single cascade of protein interactions; insulin resistance can affect any of the major sites of insulin action: the liver, skeletal muscle, or adipose tissue. The mechanism(s) by which these tissues become resistant to insulin may be numerous each with small individual effects, making them difficult to identify.
3. The case-control study design in GWA studies for T2D may inadvertently include individuals with high levels of insulin resistance in the control group, decreasing the power of the study to detect insulin resistance loci.

4. GWA studies for quantitative T2D-related phenotypes such as fasting insulin, glucose, or indices based upon them, maybe biased to the discovery of genes involved in beta cell function.

	Pre-2007		2007-2011	
	Number of Genes	**Genes**	**Number of Genes**	**Genes**
T2D	3	*PPARG, KCNJ11, TCF7L2*	50	*NOTCH2, PROX1, GCKR, THADA, BCL11A, RBMS1, IRS1, ADAMTS9, ADCY5, IGF2BP2, WFS1, ZBED3, CDKAL1, DGKB, JAZF1, GCK, KLF14, TP53INP1, SLC30A8, PTPRD, CDKN2A, CHCHD9, CDC123, HHEX, DUSP8, KCNQ1, CENTD2, MTNR1B, HMGA2, TSPAN8, HNF1A, ZFAND6, PRC1, FTO, SRR, HNF1B, DUSP9, CDCD4A, UBE2E2, GRB14, ST6GAL1, VPS26A, HMG20A, AP3S2, HNF4A, SPRY2*
BMI	1	*MC4R*	30	*NEGR1, TNNI3K, PTBP2, TMEM18, POMC, FANCL, LRP1B, CADM2, ETV5, GNPDA2, SLC39A8, HMGCR, PCSK1, ZNF608, NCR3, HMGA1, LRRN6C, TUB, BDNF, MTCH2, FAIM3, MTIF3, PRKD1, MAP2K5, FTO, SH2B1, GPRC5B, KCTD15, GIPR, TMEM160*
Glucose / Insulin	1	*GCK*	15	*GCKR, G6PC2, IGF1, ADCY5, MADD, ADRA2A, CRY2, FADS1 (MIM 606148), GLIS3, SLC2A2, PROX1, C2CD4B, DGKB, GIPR, VPS13C*
Fat distribution	0		20	*TBX15, LYPLAL1, IRS1, SPRY2, GRB14, STAB1, ADAMTS9, CPEB4, VEGFA, TFAP2B, LY86, RSPO3, NFE2L3, MSRA, ITPR2, HOXC13, NRXN3, ZNRF3, PIGC*
Total	5		107	

Table 1.4. Genes implicated for metabolic traits before and after the main wave of GWA studies (Visscher *et al.*, 2012).
This table illustrates the impact GWA studies have had on complex traits. Using metabolic traits as an example, the number of susceptibility loci identified has increased by over 2000% in just 4 years, these loci inherently implicate the biological pathways they act within, which has expanded our understanding of the molecular processes that contribute to these traits.

Chr	SNP	Nearest Gene	Beta cell function	Insulin Action
1	rs10923931	*NOTCH2*	–	–
1	rs340874	*PROX1*	–	–
2	rs780094	*GCKR*	–	Yes
2	rs11899863	*THADA*	Yes	–
2	rs243021	*BCL11A*	–	–
2	rs7578426	*IRS1*	–	Yes
3	rs13081389	*PPARG*	–	Yes
3	rs6795735	*ADAMTS9*	Yes	–
3	rs11708067	*ADCY5*	Yes	–
3	rs1470579	*IGF2BP2*	Yes	–
4	rs1801214	*WFS1*	Yes	–
5	rs4457053	*ZBED3*		–
6	rs10440833	*CDKAL1*	Yes	–
7	rs2191349	*DGKB-TMEM195*	Yes	–
7	rs849134	*JAZF1*	Yes	–
7	rs4607517	*GCK*	Yes	–
7	rs972283	*KLF14*	–	Yes
8	rs896854	*TP53INP1*	–	–
8	rs3802177	*SLC30A8*	Yes	–
9	rs10965250	*CDKN2A/B*	Yes	–
9	rs13292136	*CHCHD9*	–	–
10	rs12779790	*CAMK1D*	Yes	
10	rs5015480	*HHEX/IDE*	Yes	–
10	rs7903146	*TCF7L2*	Yes	–
11	rs2334499	Imprinted region	–	
11	rs231362/rs163184	*KCNQ1*	Yes	–
11	rs5215	*KCNJ11*	Yes	–
11	rs1552224	*CENTD2*	Yes	–
11	rs1387153	*MTNR1B*	Yes	–
12	rs1531343	*HMGA2*	–	Possible
12	rs4760790	*TSPAN8*	Yes	–
12	rs7957197	*HNF1A*	–	–
15	rs11634397	*ZFAND6*	–	–
15	rs8042680	*PRC1*	–	–
16	rs11642841	*FTO*	–	Yes
17	rs4430796	*HNF1B*	Yes	–
19	rs10423928	*GIPR*	Yes	–
X	rs5945326	*DUSP9*	–	Possible

Table 1.5. Validated T2D-implicated genes that are involved in insulin secretion or resistance (Petrie *et al.*, 2011).
The vast majority of these genes implicated in T2D susceptibility have functions that indicate a role in insulin secretion.

Mitochondria and Diabetes

All eukaryotic cells contain mitochondria that are responsible for most of the ATP-generation within the cell. In evolutionary terms, mitochondria are aerobic bacteria that formed a symbiotic relationship with primordial eukaryotic cells. Over time, the mitochondrial DNA (mtDNA) of these once free living prokaryotes have made their way into the nucleus of the eukaryotic host cell via endosymbiotic gene transfer (Timmis *et al.*, 2004). There is not much evidence for the transfer of nuclear DNA into mitochondria, presumably due to the added complexity of transcribing intron-containing eukaryotic genes and lack of ability to process the mRNA transcript within the mitochondrion. The mitochondrial genome is just 16, 571 bp in size and contains 37 genes – 22 of which are transfer RNA (tRNA) genes, 13 are protein-coding genes, and two are ribosomal RNA (rRNA) genes (Figure 1.6). The current estimate of the number of nuclear-encoded mitochondrial (NEM) genes in the human genome is 1,098 (Pagliarini *et al.*, 2008). Each mitochondrion has 2-10 copies of mtDNA and depending upon cell-type and each cell is estimated to contain 1,000 – 2,000 mitochondria. In contrast to nuclear DNA, mDNA is virtually devoid of intronic, intergenic, and repetitive sequences.

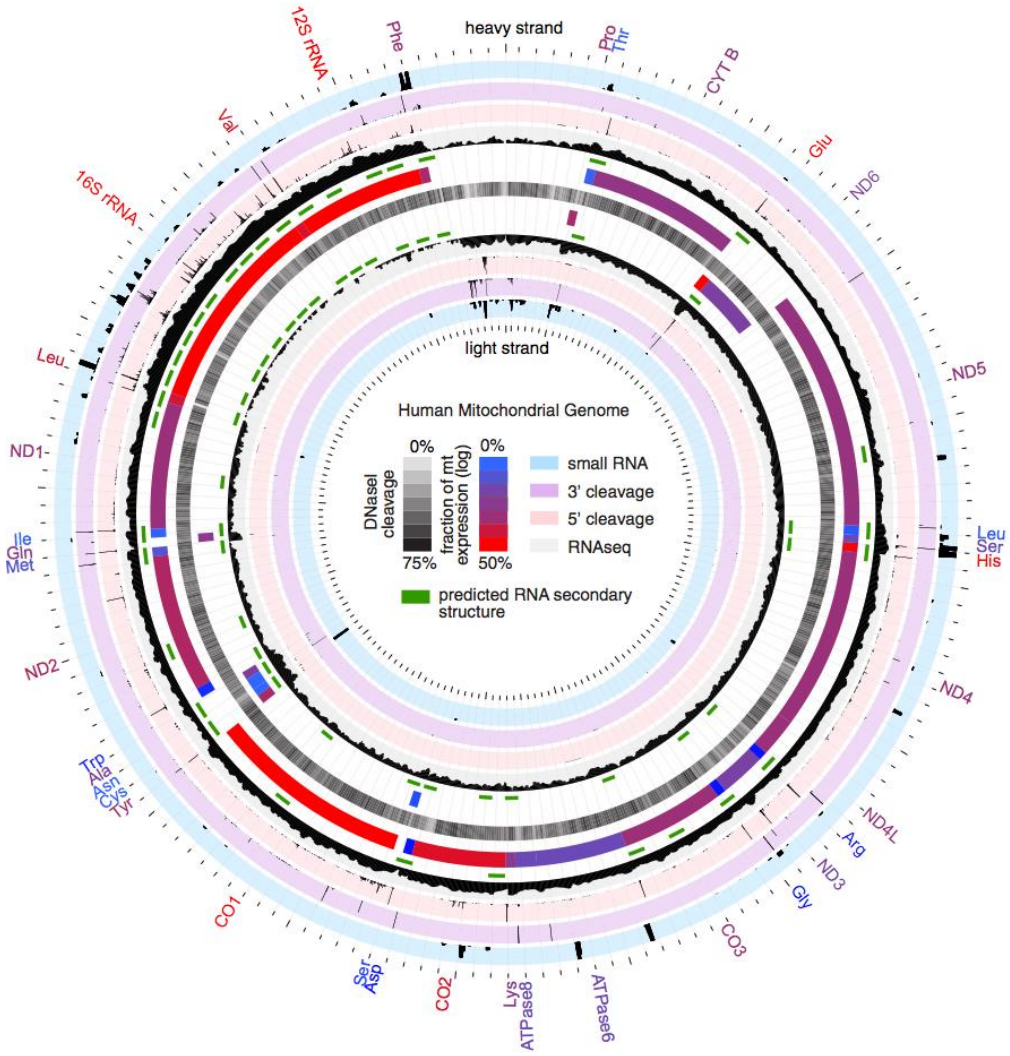


Figure 1.6. The human mitochondrial genome (Mercer *et al.*, 2011).
The human mitochondrial genome contains 37 genes, the abbreviated gene names are labelled on the outermost circle of this figure.

How do Mutations in the Mitochondrial Genome Cause Diabetes?

Maternally inherited diabetes-deafness (MIDD) syndrome (OMIM #520000) is a condition characterised by various neurological, endocrinological, and sensory features caused by mutations in one of three mitochondrial tRNA genes. Studies of patients with MIDD have shown the diabetes component of this condition resembles T1D more closely than T2D in that affected individuals have low levels of circulating insulin. A well-studied mtDNA polymorphism that causes MIDD is the A3243G mutation – an A to G transition at position 3243 in the mitochondrially-encoded tRNA (Leu, UUR) gene *MTTL1* (van den Ouweland *et al.*, 1992). This mutation causes a dimerisation in the D loop of the tRNA molecule, which destabilises the molecule and ultimately hinders mitochondrial protein synthesis and the rate of oxidative phosphorylation (Wittenhagen and Kelley, 2002). It is thought this inability to synthesise sufficient ATP via oxidative phosphorylation lowers the ATP/ADP ratio and consequently insulin is not secreted in response to glucose (Maassen *et al.*, 2004).

Contribution of Mitochondrial and Nuclear Genome Variants to T2D

The identification of a form of diabetes caused by a polymorphism in mitochondrial DNA raises the possibility that mitochondrial variants may also contribute to the common form of T2D. A recent meta-GWA scan of T2D (Voight *et al.*, 2010), revealed novel susceptibility loci for T2D were previously known to cause monogenic forms of diabetes. This overlap of loci implicated between monogenic and common T2D confirms the notion that careful genetic studies of all types of diabetes can aid our understanding of the aetiology of this clinically heterogeneous condition.

Examples of highly penetrant mutations in genes identified for monogenic and syndromic forms of diabetes that are also implicated in the common polygenic form of diabetes are presented in Table 1.6 reproduced from Vaxillaire and Froguel (2008). Some variants in these genes with minor effects on T2D susceptibly have not been detected through GWA studies, e.g. *TCF1*, *TCF2*, and *ABCC8*. This serves as a reminder that although GWA studies have shown both novel and previously-known genes to be associated with T2D, many more may be missed either due to their rarer frequency (MAF < 0.10) or small effect size (for common variants) and the high multiple penalty incurred with testing hundreds of thousands to millions of SNPs.

Gene	Monogenic Disease (OMIM)	Polygenic Type 2 Diabetes	References
HNF4A	MODY1 (125850)	Variants at the P2 promoter (MAF > 0.15). OR = 1.15 [1.00 –1.30] in Europeans	(Love-Gregory *et al.*, 2004, Silander *et al.*, 2004)
GCK	MODY2 (125851)	Variant –30G/A (β-cell promoter). OR = 1.22 [1.13–1.32] in Europeans	(Sladek *et al.*, 2007, Diabetes Genetics Initiative of Broad Institute of *et al.*, 2007)
TCF1 / *HNF1A*	MODY3 (600496)	G319S, OR = 2.0 in Oji-Cree (heterozygous carriers). OR = 1.17 [1.06–1.30] in Europeans (other variants, MAF > 0.10)	(Triggs-Raine *et al.*, 2002, Winckler *et al.*, 2005)
TCF2 / *HNF1B*	MODY5 (604284)	Intronic variants (MAF > 0.10). OR = 1.12 [1.07–1.17] in Europeans.	(Winckler *et al.*, 2007, Bonnycastle *et al.*, 2006)
TIEG2 / *KLF11*	Early-onset T2D (603301)	G62R (MAF > 0.10). OR = 1.29 [1.12–1.49] in Northern-Europeans	(Neve *et al.*, 2005)
KCNJ11	PNDM (606176)	E23K (MAF > 0.30). OR = 1.14 [1.10 –1.19] in Europeans	(Hani *et al.*, 1999, Gloyn *et al.*, 2003, Florez *et al.*, 2004, Frayling, 2007)
ABCC8	CHI, PNDM, Dominant T2D (256450)	A1369S (MAF > 0.30). OR = 1.14 [1.02–1.28] in Caucasians	(Gloyn *et al.*, 2003, Florez *et al.*, 2004)
WFS1	DIDMOAD (22233)	R611H (MAF ~ 0.60), other intronic variants. OR = 1.11 [1.08 – 1.16] in Europeans.	(Sandhu *et al.*, 2007, Frayling, 2007)

Table 1.6. Type 2 diabetes genes identified in both monogenic and polygenic forms of the disease (Vaxillaire and Froguel, 2008).
Genes that underlie monogenic and syndromic forms of diabetes also contain variants that contribute to the polygenic form of the disease. Abbreviations: MAF – minor allele frequency, OR – odds ratio.

Why Might NEM Genes be Important in T2D?

Moving from the mitochondrial genome to the nuclear genome, there is some evidence that NEM genes may play a role in T2D – for example, a nonsynonymous SNP (rs71645922, H324Q) in *LARS2* gene has been associated with T2D in 4 population samples from the Netherlands and Denmark (Hart *et al.*, 2005). *LARS2* is an interesting nuclear candidate gene as it is involved in the same pathway as the mitochondrially-encoded tRNA gene *MTTL1*, described above. A follow-up study (Reiling *et al.*, 2010) with the Diagram consortium (Zeggini *et al.*, 2008) and individuals from several European samples found no significant association despite the large sample size (n = 35,712).

A search of the National Human Genome Research Institute (NHGRI) catalogue of published GWA Studies (Hindorff *et al.*, 2009) for T2D and related traits show six NEM genes (Pagliarini *et al.*, 2008) to be associated with these traits. These six genes are *MRPL33*, *FITM2*, *TRIAP1*, *C6orf57*, *HK1*, and *IDE*, with mitochondrial functions that range from ribosome components to degradation of mitochondrial targeting sequences.

Rather than taking a hypothesis-free approach to identifying T2D susceptibly variants in NEM genes, some studies have looked specifically at NEM genes that encode the proteins of the oxidative phosphorylation (OXPHOS) pathway. This pathway is central to T2D as the production of ATP is one of the key signals in the pancreatic beta cell for glucose-stimulated insulin secretion (Henquin, 2009) and avoidance of hyperglycaemia hinges on insulin secretion overcoming insulin resistance. Moreover, by only looking at variants in genes in the OXPHOS pathway and not the entire genome, it lessens the penalty incurred by multiple testing.

Another study (Reiling *et al.*, 2009) has looked at the role of 13 "essential" NEM genes involved in mitochondrial transcription, translation or replication in order to test for association with T2D. In their two-stage study design, they first analysed 58 SNPs, which tagged nearly all common SNPs in 13 genes, for association with T2D. The second stage took the top SNPs (minimum p-value = 0.005) from stage one, genotyped and tested them for association with T2D in three additional Dutch cohorts

for replication. None of the nominally associated SNPs in the initial stage were replicated.

The majority of proteins in the OXPHOS pathway are encoded by nuclear genes (Baughman *et al.*, 2009). A study (Olsson *et al.*, 2011a) attempted to identify any common SNPs in or near OXPHOS genes using clinical phenotypic data from the Diabetes Genetics Initiative (DGI) GWA scan for T2D. Nine SNPs from six genes showed nominal association with T2D with p-values less than 1×10^{-2} in the initial DGI GWA scan, of these, the two SNPs with lowest p-values (rs606164, p-value = 9×10^{-4}; rs1323070, p-value = 3×10^{-3}) were selected to be followed up with *in vivo* studies. Both SNPs were nominally associated (p-value < 0.05) with an insulinogenic index (following an oral glucose tolerance test: [[insulin at 30 minutes – insulin at 0 minutes] / [glucose at 30 minutes – glucose at 0 minutes]]).

In a similar study design, Snogdal *et al.* (2012) investigated 10 variants in six OXPHOS genes that showed nominal association (p-value < 1×10^{-2}) in a T2D GWAS meta-analysis (Zeggini *et al.*, 2008) by attempting to replicate their associations in an independent Danish cohort. Only one of the ten SNPs reached a nominal association with T2D (rs9915302, odds ratio, OR = 1.14 [1.02 – 1.28], p-value = 0.02) in the Danish cohort. Combining their Danish cohort with a second T2D meta-GWAS (Voight *et al.*, 2010), improved the number of nominal associations to six. The authors also looked at associations with diabetes-related phenotypes, again only nominal associations (p-value ≈ 0.05) were found. None of the original 10 SNPs were significantly associated with any phenotype after correction for multiple testing in the three analyses in the study (Snogdal *et al.*, 2012).

A problem with these studies (Olsson *et al.*, 2011a, Snogdal *et al.*, 2012) is that they lack any examination of the extent of LD surrounding the selected SNPs, limiting themselves to the initial GWA scan results. It is known the causal variant(s) is rarely, if at all, identified via a GWA scan. The hope in performing a GWA scan is that the casual variant(s) lies within a haplotype block that also harbours a genotyped SNP. The goal of GWA scan is to elucidate the genetic pathways underlying phenotypic variation, further work is then needed to identify the causal genetic variant(s) within

the loci identified by the GWA scan that contribute to phenotypic variation. Therefore to focus on purely on SNPs, that did not even reach genome-wide significance, then to show inconsistent or weak association in separate, albeit large samples, the findings (Olsson *et al.*, 2011a, Snogdal *et al.*, 2012) are not convincing and naïve to conclude that NEM genes to do not contribute to T2D.

Reasoning that as previous studies looking at candidate NEM genes failed to detect any robust associations with T2D or related phenotypes, it has been suggested (Segre *et al.*, 2010) that the effect of an individual NEM gene variant is too small to be detected and that gene-set based pathway analyses may be more fruitful. To that end, a modified Gene Set Enrichment Analysis, GSEA (Subramanian *et al.*, 2005) approach called Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) has been developed by Segre *et al.* (2010). While GSEA was developed to analyse gene expression data, MAGENTA was developed to analyse SNP p-values from a GWA scan. Briefly, the MAGENTA pipeline involves mapping SNPs to genes, assigning the minimum SNP p-value to the gene, adjusting for gene size, number of SNPs, LD per kilobase (Kb), and calculating an association enrichment score for each gene set over and above that expected under the null hypothesis. The enrichment score is determined as the fraction of genes in a gene set with a p-value less than a predetermined threshold and the gene set p-value (or evidence of enrichment) is obtained by comparing this fraction with a fraction obtained from a random gene set of identical size.

This MAGENTA method was applied to T2D GWA data to test for enrichment of NEM gene regulators of mitochondrial gene transcription, NEM genes pertinent to OXPHOS, and all known NEM genes. No enrichment was found for T2D or related phenotypes, which failed to support the original application of GSEA (Mootha *et al.*, 2003) that found enrichment for the down regulation of a subset of OXPHOS genes in skeletal muscle from individuals with IGT or T2D. A likely reason for why Segre *et al.* (2010) did not observe results similar to Mootha *et al.* (2003) is that gene expression data lends itself better to gene-centred analyses than SNP data does.

It is common in GSEA-type analyses of GWA data to use the minimum SNP p-value of each gene (Wang *et al.*, 2007, Torkamani *et al.*, 2008, Askland *et al.*, 2009) but it is

a poor measure of the importance of each gene in a pathway. This form of reductionism disregards the presence of multiple independent signals with a gene and fails to adequately control for gene size or degree of linkage disequilibria. It would be more desirable to incorporate all SNP p-values such as methods described by Purcell *et al.* (2007), which takes the mean association value for all SNPs within a gene, or O'Dushlaine *et al.* (2011), which computes a ratio of significant and non-significant SNPs for each gene.

Research Aims of this Thesis

The three main research aims of this thesis are:

1a. The validation and estimation of an accurate proxy measure for visceral fat accumulation using DXA and simple anthropometrical measures.

1b. Investigate whether visceral fat mediates the observed association between different measures of body adiposity and common cardiovascular and metabolic morbidities, including T2D.

2. Determine if the locus encompassing two genes encoding the presenilin-associated, rhomboid-like (*PARL*) and the ATP-binding cassette, sub-family C (CFTR/MRP), member 5 (*ABCC5*) proteins is a novel disease susceptibility locus for T2D.

3. To identify NEM pathways associated with T2D in European and African American samples.

Chapter 2: Materials and Methods

Samples

TwinsUK Sample

Subjects used for the studies in this thesis include adult twin volunteers from the Department of Twin Research and Genetic Epidemiology (DTR), which is based at St. Thomas' Hospital in London, United Kingdom (Spector and Williams, 2006, Moayyeri *et al.*, 2013, Moayyeri *et al.*, 2012). Recruitment to the registry (TwinsUK) is largely through campaigns in the national press, TV, radio, through their own website (www.twinresearch.ac.uk) and recently social networking websites. The registry currently consists of 12,000 same-sexed twins - approximately 80% are female and the ratio of monozygotic (MZ) and dizygotic (DZ) twins is roughly 1:1. Twins have participated in a variety of studies over the last two decades; there have been three major phases of TwinsUK (Figure 2.1): the baseline phase ran from the registry's inception in 1992 until 2004, the first follow-up phase took place between 2004-2007, whilst the second follow-up, known as the Healthy Ageing Twin Study (HATS), ran between 2007-2010. The aim of the HATS was to provide longitudinal data for female twins over 40 years old and had previously visited the hospital.

During these phases, many types of studies were conducted; these included questionnaire-based, posted collection kits and clinical visits at the hospital. At the clinical visit, in addition to extensive health questionnaires, routine biological samples were collected and clinical assessments were performed during the visit to the hospital. For specialised studies, such as those focused on eye-related phenotypes, additional study-specific phenotypes were collected, therefore the number of phenotypes differs among twins. A comprehensive list of phenotypes collected can be downloaded from http://www.twinsuk.ac.uk/data-access/phenotypes.

Upon joining the registry all twins provided informed written consent for proposed usage and storage of their information and biological specimen(s). Each individual is assigned a unique study number identifier to ensure anonymity. For each subsequent round of study participation, twins are re-consented, which ensures the individual understands the purpose of the study and is given the opportunity to withdraw from the study. St. Thomas' Hospital Research Ethics Committee approved all studies conducted by the DTR.

Data used in this study were originally collected and managed by staff and PhD students of the DTR, including myself for the first two years of my PhD. However, I did not collect the genotypic, DXA and glycaemic data that are central to some of the questions addressed in this thesis.

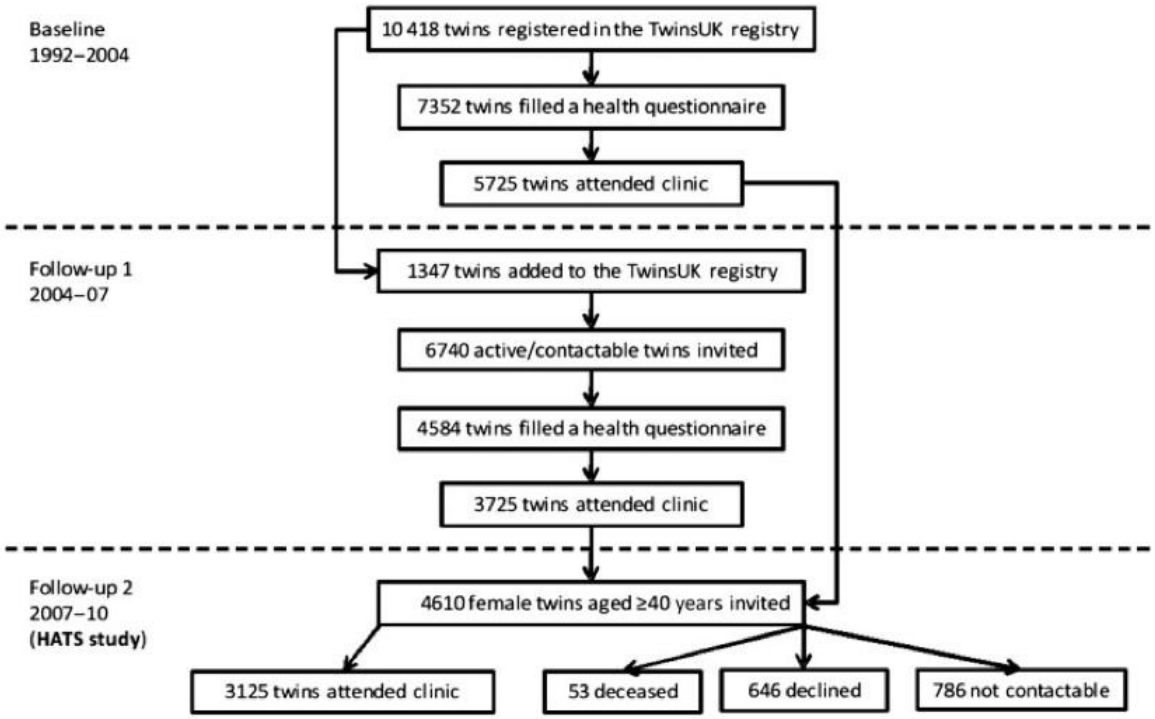


Figure 2.1. Schematic of the major TwinsUK recruitment phases (Moayyeri *et al.*, 2012).
Participation levels for the TwinsUK volunteers from 1992-2010; data analysis in this thesis only used data from the most recent data clinical visit if more than one was recorded.

As the focus of the thesis is T2D and related glycaemic traits, the main biological phenotypes used are fasting plasma insulin and glucose. Twins were asked to fast for at least 8 hours preceding their visit to St Thomas' Hospital. Visits took place both in the morning and afternoon; unfortunately the time of visit had only recently (post-2007) been recorded. Insulin and glucose measurements by the DTR since its formation are listed in Table 2.1. The nurses and research assistants of the DTR performed collection of blood.

Dual-Energy X-ray Absorptiometry

Research assistants at the Department of Twin Research performed whole body dual-energy X-ray absorptiometry (DXA) scans in the supine position, using a fan beam X-ray bone densitometer (QDR-4500W, Hologic, Massachusetts) and analysed with QDR System Software Version 12.6 (Hologic, Inc., Massachusetts). The DXA scanner was calibrated daily with a spine phantom and weekly using the step phantom as instructed by the manufacturer's guidelines.

Computed Tomography

Computed tomography (CT) scans used in this thesis were provided by Dr Marina Cecelja. Scan details are as follows: subjects were placed in the supine position and a truncal CT scan was performed using a CT helical scanner (Brilliance CT component of the Philips Precedence SPECT/CT system, Philips, Eindhoven, Netherlands) with slice thickness set at 5 mm. The total cross sectional area (CSA) was intersected at the level of the intervertebral disc between lumbar vertebrae four and five (L4:L5). CT scans were analysed using Osirix X imaging software, version 3.7.1 (Pixmeo, Switzerland) to provide X-ray attenuation data at the pixel level.

MuTHER Fat Expression

During the clinical visits to St Thomas' Hospital, approximately 850 twins had abdominal skin punch biopsies performed to extract skin and subcutaneous fat, as well as a blood sample to generate lymphoblastoid cell lines (LCLs). Details of the biopsy method can be found elsewhere (Nica *et al.*, 2011). Genome-wide messenger RNA expression (~48,000 probes) was measured in these three tissues by the Multiple Tissue Human Expression Resource (MuTHER) project (http://www.muther.ac.uk/) using the Illumina expression array HumanHT-12 version 3 array.

Comprehensive details of the data normalisation process are described in the original reports (Nica *et al.*, 2011, Stranger *et al.*, 2007); in summary data generated for each tissue was first transformed on the $\log_2$ scale, then quantile normalised across individual replicates then across all individuals. The normalised fat expression data used in Chapter 4 is now publically available (http://www.muther.ac.uk/Data.html).

	Duration	Assay	Instrument / Analyser	Reference range	Original units
Glucose	1992-August 2000	Enzymatic rate / Colorimetric assay	Kodak Ektachem dry chemistry analysers (Johnson and Johnson Vitros Ektachem machine)	3.5 - 5.5	mmol/L
	August 2000-March 2006	Time rate / Conductivity Glucose concentration is determined by the oxygen rate method employing a Beckman Oxygen electrode. Electronic circuits determine the rate of oxygen consumption, which is directly proportional to the concentration.	Beckman LX20 analysers	3.5 - 5.5	mmol/L
	April 2006 - present	Enzymatic rate / Colorimetric assay	Roche P800 modular system	3.5 - 5.5	mmol/L
Insulin	1992 - 2000	Microparticle Enzyme Immunoassay (Abbott Laboratories Diagnostics Division, Maidenhead, U.K.)	Abbot IMX Analyser	0 - 15.6	µU /ml
	1996 - May 2007	Solid-phase two-site chemiluminescent assay	Immulite 100 analyzer (DCP Ltd)	8.9 - 28.4	µU /ml
	2007 - present	Chemiluminescence	Roche 2010 analyser	17.8 - 173	pmol/L

Table 2.1. Assay and instrument details for measuring insulin and glucose from 1992-2012 at St Thomas' Hospital, London, UK.
Various instruments were used to measure insulin and glucose in the TwinsUK sample. The duration each assay was in use for was useful in handling potential batch effects, as was the manufacturers' reference range.

Wellcome Trust Case Control Consortium

Established in 2005, the Wellcome Trust Case Control Consortium (WTCCC; http://www.wtccc.org.uk) comprises of more than 50 British research groups with the aim of investigating the genetics of disease susceptibility. Phase one of the project (WTCCC1) looked at seven diseases (T1D, T2D, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis and Crohn's disease) in comparison with an ostensibly healthy shared control group selected from the 1958 Birth Cohort and the UK Blood Services Controls (UKBS).

Sample recruitment and phenotype definition for each of the seven diseases is fully described in their original report (Wellcome Trust Case Control, 2007). T2D cases were identified from the Diabetes UK Warren 2 repository, diagnosis was determined if the individual has been prescribed treatment with hypoglycaemic agents and/or previous evidence of hyperglycaemia. T1D, MODY, and mitochondrial diabetes cases were excluded based on personal and family history; current or previous gestational diabetes was not reported in the exclusion criteria. In the T2D GWA scan, three SNPs (rs9465871, rs4506565, rs9939609) were associated with T2D, all of which were in or near genes that had previously been implicated by other GWA studies.

National Institute for Diabetes and Digestive and Kidney diseases (NIDDK)

The NIDDK focuses on diabetes (all types) and common complications arising from it (Fradkin and Rodgers, 2013). It was formed in 1950 in Maryland (USA) and is funded by the National Institutes of Health (NIH). At the beginning of 2012, members of the NIDDK reported a GWA study of T2D in African Americans (Palmer et al., 2012); cases were adults over 25 years of age with T2D and end stage renal disease (ESRD, discovery cohort) and T2D without ESRD for the replication phase, recruited from health centres in North Carolina and neighbouring states. Controls were individuals recruited from community clinics in the same area as the cases.

Genotyping

Genotype data used throughout this thesis originate from a number of different platforms, including fine map genotyping undertaken specifically for studies presented in this thesis. Here I outline the key attributes of each starting with the TwinsUK, then the publically available WTCCC1 and the NIDDK African American GWA resources obtained from dbGAP.

TwinsUK

TwinsUK have genome-wide SNP data for 5, 701 individuals split across the Illumina HumanHap 300K and the Illumina HumanHap 610 Quad platforms. Staff from the DTR took blood from the twins and genotyping was performed on DNA extracted from whole blood at the Wellcome Trust Sanger Institute (Cambridgeshire, United Kingdom) and the Center for Inherited Disease Research (CIDR; Maryland, United States of America).

The Illumina HumanHap300 interrogates approximately 317,000 SNPs, estimated to capture around 80% of SNPs with MAF > 5% in HapMap (European ancestry; CEU) phases 1 and 2 (Illumina HumanHap300 data sheet). Markers are evenly distributed throughout on a physical map scale with an average of one SNP per nine Kb. The Illumina Human610-Quad platform has around 620,00 SNPs, with coverage of the HapMap (CEU) release 23 data (four million SNPs) by this platform estimated to 89% and the mean physical spacing between markers on the platform is one SNP per five Kb (Illumina HD analysis data sheet).

For the HumanHap300, 2,226 samples were genotyped, including 27 duplicates, thus 2,199 unique individuals; for the Human610-Quad platform, 3,512 samples were genotyped, 10 of which duplicate samples. In terms of zygosity, the 2,199 individuals on the HumanHap300 platform comprised of 1,416 DZ, 780 MZ, and three with unknown zygosity at the time of genotyping; the 3,502 individuals on the Human610-Quad platform comprised of 1,977 DZ, 1,485 MZ, and 40 with unknown zygosity at the time of genotyping.

The SNP data from both platforms were checked for quality control (QC) by the staff of the DTR, also described in Richards *et al.* (2008); in summary this involved filtering out genotype calls based on the following: a genotype calling posterior probability less than 95%, a call rate less than 90%, deviation from Hardy-Weinberg equilibrium (p value < $1.0x10^{-4}$), or minor allele frequency (MAF) less than 1%. The sample genotyped on the HumanHap300 platform has previously been subjected to population stratification analysis implemented in Structure (Pritchard *et al.*, 2000) and showed the TwinsUK sample to be homogenous.

For the work presented in Chapter 4, the HumanHap300 and Human610-Quad samples were initially analysed separately due to different marker coverage and also to check for potential phenotypic heterogeneity with respect to fasting plasma insulin and glucose. Imputation was not used to increase marker coverage in the thesis because some of the analyses in Chapters 4 and 5 require empirical modelling of all marker pair-wise relationships.

WTCCC1

In total, 17,000 samples were genotyped on the Affymetrix 500K platform for WTCCC1; individual level genotype data can be downloaded from the European Genotype Archive (http://www.ebi.ac.uk/ega). The genotyping assay protocol is described in the WTCCC1 initial report (2007); coverage is substantially less than the TwinsUK 610-Quad platform with 392,575 SNPs passing QC and MAF > 1%. The coverage on the Affymetrix 500K is more evenly distributed than the HumanHap 300K, with the latter being more targeted towards SNPs within exons (Table 2.2).

	HumanHap 300K	**Affymetrix 500K**
Intragenic (between Tx start and stop)	123,373 (0.39)	184,912 (0.37)
Exonic	12,023	8,513
Intronic	111,350	176,399
Within 10Kb of Tx start or stop	35,290 (0.11)	52,208 (0.10)
Intergenic	158,839 (0.50)	236,706 (0.53)

Table 2.2. Comparison of SNP coverage between the HumanHap 300K and Affymetrix 500K (DeWan *et al.*, 2007).

Number of SNPs given in table with proportion of total SNPs given in parentheses. Abbreviation: Tx – transcriptional.

African American

Phenotype and genotype data for the NIDDK African American sample were obtained from the National Center for Biotechnology Information (NCBI) database of genotypes and phenotypes, dbGaP (http://www.ncbi.nlm.nih.gov/gap). The original genotyping and QC details can be found in Palmer *et al.* (2012); here I only note that post-QC coverage on the Affymetrix Genome-wide Human SNP array 6.0 platform exceeds 800K SNPs, which is greater than both the TwinsUK Human610-Quad (583K post-QC SNPs) and the WTCCC1 Affymetrix 500K (393K post-QC SNPs) platforms.

One major difference between TwinsUK (Illumina) and the publically available data sets (Affymetrix) is the genotyping method. The Illumina HumanHap300 and the 610-platforms utilise the single-base extension genotyping method (Steemers *et al.*, 2006). In short, an oligonucleotide 50 bp in length is hybridised adjacent to the SNP, labelled dideoxynucleotides are incorporated at the SNP locus, which identifies the allele at the locus. The Affymetrix genotyping system was used for the WTCCC1 (Affymetrix 500K) and the NIDDK African American (Affymetrix 6.0) GWA studies; the Affymetrix system requires a complex reduction stage prior to differential hybridisation of labelled DNA to the quartet of perfect and mismatched pairs of 25-mer probes on the array.

Quantitative Genetic Theory

Biometrical genetics is the study of the inheritance of a continuous phenotype. A phenotypic measurement is the outcome of latent genetic and environmental contributions:

$$P = G + E$$

Where P is the phenotypic mean value, G is the genotypic value, and E is the environmental deviation. The focus of the following section is the partitioning of genotypic contribution into additive and dominance components:

$$G_{total} = G_{additive} + G_{dominance}$$

Genotypic values can be assigned to specify the phenotypic value for each genotype. If one initially assumes that there is no intra-locus dominance effect at an autosomal biallelic locus with alleles A_1 and A_2, whereby A_1 increases the phenotypic value, the genotypic effect of the homozygote A_2A_2 and A_1A_1 genotypes are -a and +a, respectively. In this example the alleles act additively in that the genotypic effect of heterozygote, A_1A_2, is exactly midway between -a and +a.

Any deviation from this additivity is termed the degree of dominance (d), which can range from none to complete. If allele A_1 is dominant over A_2, d is greater than zero, and conversely, if allele A_2 is dominant over A_1, d is less than zero (Figure 2.2). For complete dominance, d = -a or +a.

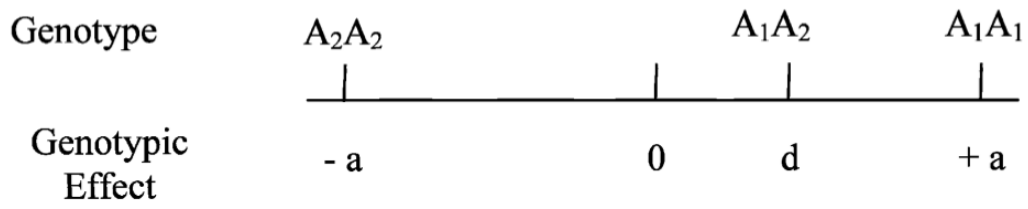


Figure 2.2. Illustration of dominance at a biallelic locus (Evans *et al.*, 2002).

As genotypes are not passed from parents to offspring, the allelic effect must be considered instead of the genotypic effect. This new measurement, the average effect (α) of an allele, combines allele frequencies with the genotypic effect at a locus: $\alpha = a + d(q - p)$, where p and q represent the allele frequencies for A_1 and A_2 respectively.

The sum of the average effect of the alleles at each locus an individual possess is its breeding value, also known as the additive genetic value. For example, for a single locus (A) the breeding value of an A_1A_1 homozygote is twice αA_1, for an A_1A_2 heterozygote it is the sum of $\alpha A_1$and αA_2, and lastly for the A_2A_2 homozygote, the breeding value is twice αA_2. This the mean genotypic value at the A locus in the offspring as it is the sum of the breeding values of its parents and thus individual breeding values is important in designing breeding programmes for a trait.

In the derivation of the additive genetic value of an individual, the effects of dominance and interaction between alleles at different loci (epistasis) are initially ignored. As noted earlier, the genotypic value of a locus is attributable to additive genetic effects, while deviations from additivity can be attributed in part to dominant and epistasis genetic effects. For two or more loci, we take the sum of the additive genetic and dominance values at the loci and include a term for the any interaction between loci.

$$G_{total} = G_{additive} + G_{dominance} + G_{interaction}$$

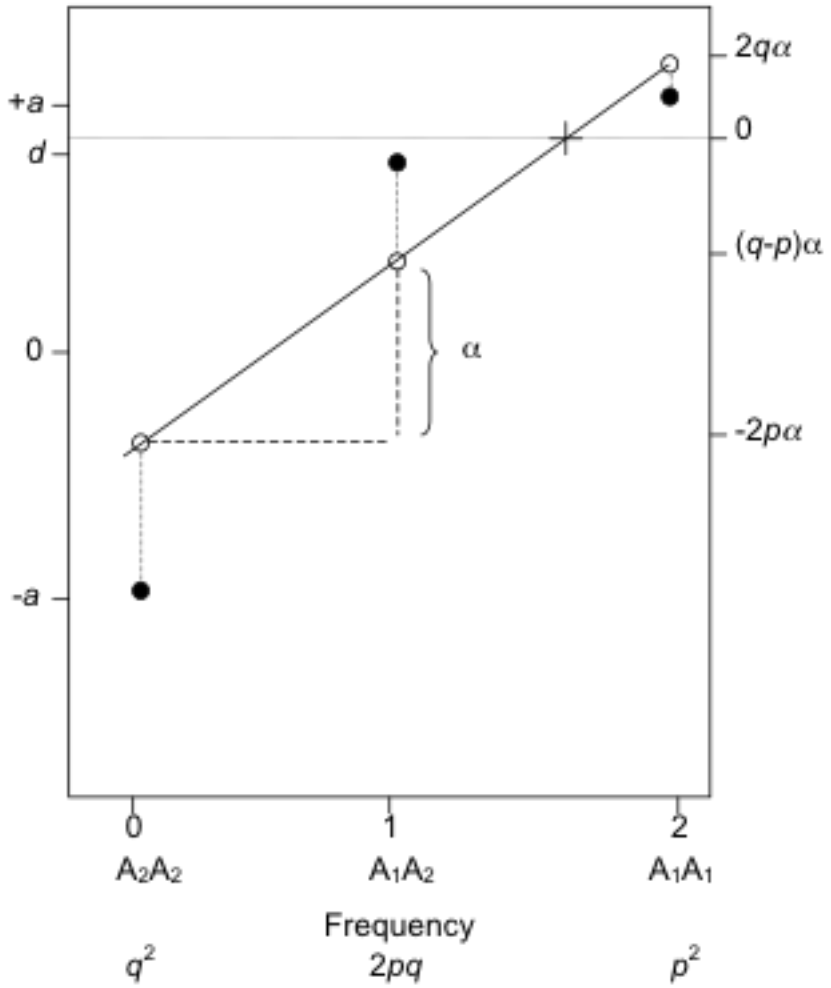


Figure 2.3. The relationship between genotypic value, genotype frequency, and additive genetic values (Falconer, 1989).

The genotypic value (open circles, y1-axis) is plotted against genotype frequency (x-axis) and the additive genetic value (closed circles, y2-axis). The A_1 allele is dominant over the A_2 allele to the extent that the genotypic value for the A_1A_2 heterozygote is almost equal to the genotypic value of the A_1A_1 homozygote. The average effect (α) is indicated between the A_2A_2 homozygote and the A_1A_2 heterozygote, but is identical in magnitude to the value between the A_1A_1 homozygote and the A_1A_2 heterozygote.

Genetic Variance

The focus of the previous section was the derivation of the breeding mean value of an individual; here I will describe the second order moment of the breeding value, the variance. The partitioning of breeding value variance is central to estimating heritability.

Total phenotypic variance (V_P) comprises of genetic (V_G) and environmental (V_E) components of variance:

$$V_P = V_G + V_E$$

The biometric model tells us that the genetic component is a culmination of additive genetic effects, with deviations from additivity in the form of dominance and interaction terms between alleles at different loci (epistatsis). Therefore, the total phenotypic variance in the can be expressed as:

$$V_P = V_A + V_D + V_I + V_E$$

The ratio of genetic variance to observed phenotypic variance is the broad sense heritability (H^2) of that phenotype. Replacing the total genetic variance with just the additive genetic variance, i.e. what is transmitted from parents to their offspring, in the ratio gives the narrow-sense heritability (h^2).

Having described how the genetic component of phenotypic variance can be further decomposed into additive and dominance component, it is important to note that a relationship may exist between the genotype and the environment. For example a genotype maybe correlated with an environment – i.e. if nutrition of an animal was determined by its phenotype e.g. weight, there is a correlation between phenotype and environment. As a portion of the individual's phenotype is determined by the sum of its genotypes, there is a correlation between genotype and environment.

In human genetics these considerations of correlation and interaction between genotypes and the environment are often assumed to be absent. The reason for assuming an absence of genotype-environment correlation is that testing for such correlation requires study designs that are difficult to implement such as inbreeding and/or complete environmental control. Genotype-environment interactions are also ignored because the prerequisite for testing this relationship is that the environmental factor contributing to the interaction is known and accurately measured.

Genetic Model Specification

The true underlying genetic model of a risk allele is seldom known. There are a few key models that describe the effect of the risk genotype on the phenotype – additive, multiplicative, codominant, dominant, and recessive. A starting point in genetic tests of association in case-control data is a count of observed genotypes at a SNP in the case and control groups (the full data box in Figure 2.4), the test is Pearson's χ^2, which is described in a following section (Allelic and Genotypic Tests of Association). This is the safest type of analysis to perform as it does not assume a genetic model, thereby avoiding specifying the wrong genetic model, which leads to a reduction in statistical power (Lettre *et al.*, 2007), Figure 2.5. It might be attractive to test all genetic models to determine the model that produces the largest test statistic. However this would incur both a multiple testing penalty and increase the computational burden by four, which is substantial on a genome-wide scale with approximately a million SNPs to be tested.

Where there is prior evidence to suggest a particular genetic model (for example, Leu262Val in Chapter 4) it is best to fit a genetic model rather than the generic Pearson's χ^2 test. The models vary slightly in their interpretation depending on the data type, i.e. quantitative trait or case-control status. An additive model is one where the disease risk increases with increasing number of risk alleles, so that heterozygotes will have a phenotypic value or risk of case status approximately midway between the values for the two homozygotes. This trend can be tested using the Cochran-Armitage test for trend. A multiplicative model tests allele rather than genotype counts and the presumed risk of disease increases by a risk factor "r" with number of risk alleles, e.g. if "B" is the risk allele, the risk for AB heterozygotes is r and for BB homozygotes is

r^2. Note the comparison between the risk genotype in additive and multiplicative models, which are 2r and r^2, respectively. If the disease risk conferred by the heterozygous genotype lies between the two homozygous genotypes but not approximately midway (additive model), then the model is said to be codominant.

The dominant and recessive models work by collapsing the three observable genotypes at a biallelic SNP into just two groups; this example is illustrated in Figure 2.4 (Lewis and Knight, 2012). In the dominant model, assuming the "T" allele is the dominant allele, the phenotypic value for GT heterozygotes will be the same as the TT homozygotes. Similarly, in the recessive model, the GG homozygotes and GT heterozygotes are collapsed into one group.

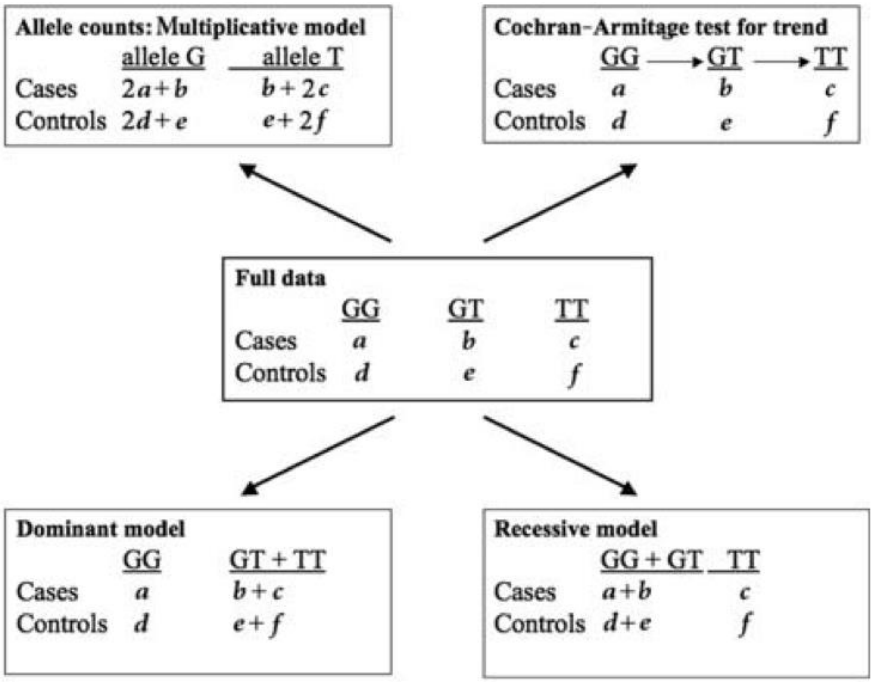


Figure 2.4. Genetic models to test for SNP association in case-control data (Lewis and Knight, 2012).
Terms "a"-"f" indicate observed genotype or allele counts of the genotype in the cases and controls.

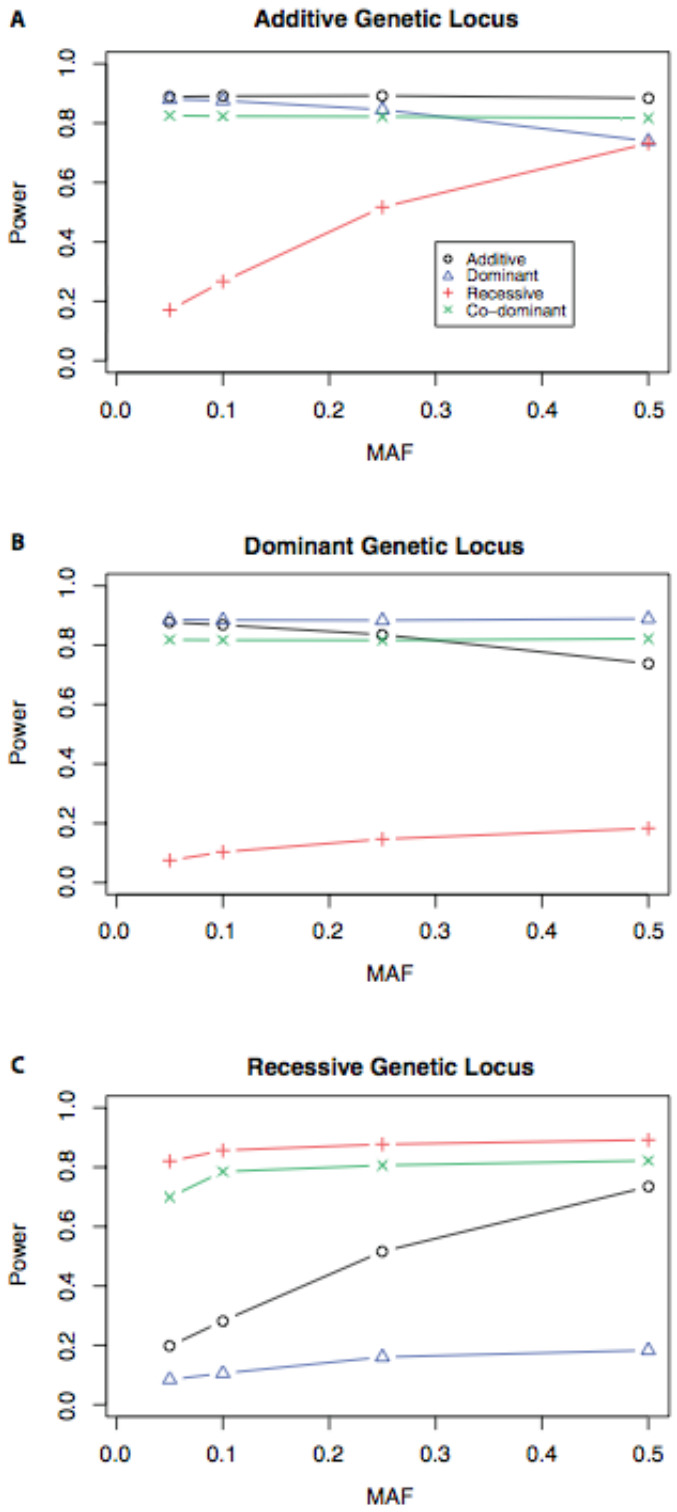


Figure 2.5. Effect of specifying inappropriate genetic models on statistical power (Lettre *et al.*, 2007).
Plots show the effect on power by fitting additive, dominant, recessive and codominant statistical models to additive, dominant, and recessive loci. Abbreviations: MAF – minor allele frequency.

Ordinary Least Squares Regression

If the relationship between a quantitative trait and SNP is linear, a simple ordinary least squares (OLS) linear regression can be used to test for association. For example, if one assumes a biallelic SNP with three genotypes – AA, Aa, and aa, each genotype is considered by the number of minor alleles it possess. The minor allele of a SNP, as the name suggests, refers to the less frequent allele at that SNP, thus if the "A" allele is more common than the "a" allele in a sample, the genotype allele counts are as follows: AA – 0, Aa – 1, and aa – 2. The OLS regression model is described as:

$$y = c + \beta_{SNP} + e$$

Where,

- y is the estimated quantitative phenotype
- c is the intercept term, i.e. the fitted value of the phenotype for the homozygous common genotype (0).
- β_{SNP} represents the beta coefficient of the SNP, which is the slope of the regression line (Figure 2.6)
- e is the residual error term and represents the difference between the actual values of the quantitative trait and the fitted values based on the regression model.

Such a model is also known as a linear dose response model, for each additional copy of the "a" allele, the estimated trait value will increase by the value of β_{SNP}. The intercept term, a, is interpreted as the trait estimated trait value for individuals with no copies of the "a" allele.

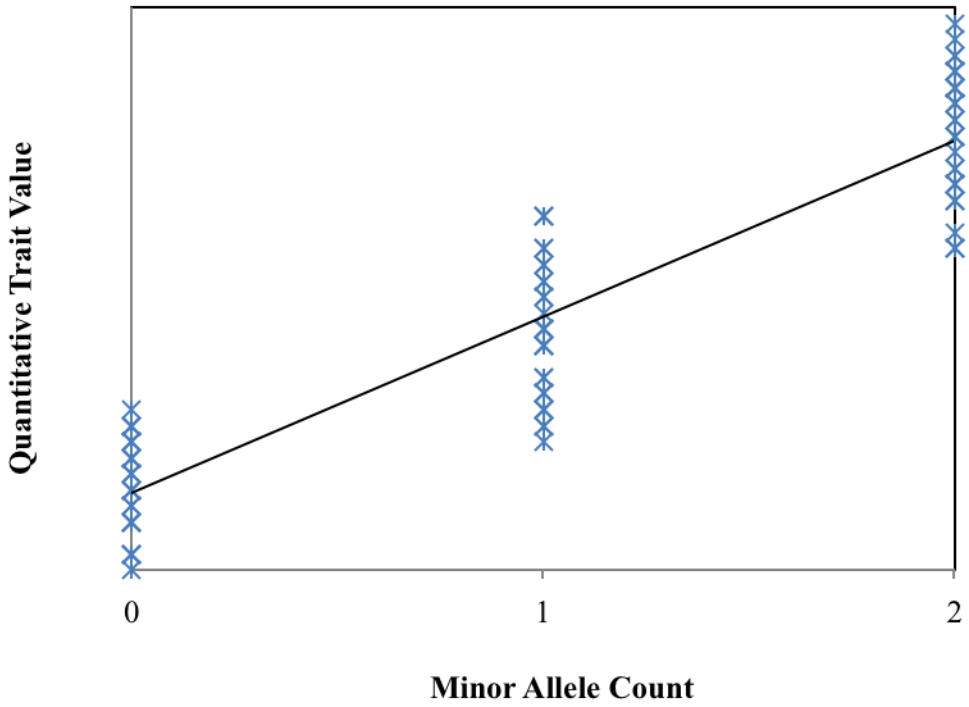


Figure 2.6. Ordinary least squares regression of quantitative trait and SNP minor allele count.

Covariates can also be added to the model that account for confounders or additional explanatory variables, for example if we assume age is associated with fasting insulin level and want to test if a SNP was also associated, the model would become:

$$y = c + \beta_{SNP} + \beta_{age} + e$$

$\beta_{SNP} + \beta_{age}$ are partial coefficients – the partial coefficient for the SNP is the effect increasing the SNP variable by one whilst adjusting for the influence of age on the phenotype.

Statistical software will often report descriptive statistics of the model such as the standard error, a test statistic, the p-value of that test statistic for each beta coefficient. The standard error of the beta coefficient (SEβ) is an estimate of the extent to which the sample beta coefficient varies about the true population mean. The test statistic, t, is simply the beta coefficient divided by its standard error and follows a t distribution, which is asymptomatically normal for large sample sizes. The significance of the t statistic is assessed by comparing its value to a critical value for a particular

significance threshold (e.g. 0.05, 0.01, 0.001, etc) and degrees of freedom. If the p-value for the test statistic is below a significance threshold, then there is evidence to reject the null hypothesis (H_0) that β_{SNP} is equal to 0 and the result favours the alternative hypothesis (H_1) that β_{SNP} is not equal to 0.

An explanatory variable might be statistically significant and therefore associated with the response variable, but still might not explain much of the variation in the response. The proportion of variance of the response variable explained by the model is given by coefficient of determination (R^2). For a multiple regression, R^2 can be adjusted by the number of explanatory variables included in the model.

Logistic Regression

If the phenotype is binary, such as disease status, which can be presented as unaffected (0) or affected (1), association between the binary phenotype and SNP (or any continuous phenotype) can be tested using logistic regression or contingency tables. In comparison to OLS linear regression described above, a logit transformation is applied to the outcome variable to convert the binary variable into a continuous variable.

$$\text{Logit}(p) = \ln (p / 1\text{-}p) = c + \beta_{SNP}$$

Where,

- ln is the natural logarithm
- p is the affected proportion
- 1-p is the unaffected proportion

If the explanatory variable is a SNP coded as 0, 1, or 2 with increasing count of the minor allele as above, the interpretation of the beta coefficient from this regression is the increase in risk for each additional minor allele. The beta coefficient can also be expressed as an OR by taking the exponential of the beta coefficient.

Many statistical software packages calculate a pseudo-R^2 model-fit statistic that is analogous (but not directly comparable) to the OLS regression R^2. While R^2 can be interpreted as the proportion of variance explained by the model, pseudo-R^2 is loosely interpreted as the proportion of variation in risk liability explained by the model (StatCorp, Texas).

Allelic and Genotypic Tests of Association

Tests of association can be performed on allele or genotype counts. For a biallelic test (Table 2.3), the two alleles of the SNP are summed across the affected and unaffected groups, compared to their expected counts and a Pearson's χ^2 statistic is calculated. The null hypothesis being tested is no association between the alleles and the affection status, it is a one degree of freedom test as degrees of freedom = (number of rows−1)(number of colums−1) = (2-1)(2-1) = 1.

The alternative test is a genotypic association (Table 2.4), which works on the same principle, the difference being that, rather than alleles, the number of genotypes at a single locus are counted for the affected and unaffected groups. As three genotypes are possible at a biallelic locus, the degrees of freedom for this test can be two (equivalent to testing for additive and dominant terms) or one (equivalent to an additive test of association).

	B	b	Total
Cases	B_{cases}	b_{cases}	$B_{cases} + b_{cases}$
Controls	$B_{controls}$	$b_{controls}$	$B_{controls+} b_{controls}$
Total	$B_{cases} + B_{controls}$	$b_{cases} + b_{controls}$	N

Table 2.3. Allele contingency table.

	bb	Bb	BB	Total
Cases	bb_{cases}	Bb_{cases}	BB_{cases}	$bb_{cases} + Bb_{cases} + BB_{cases}$
Controls	$bb_{controls}$	$Bb_{controls}$	$BB_{controls}$	$bb_{controls+} Bb_{controls} + BB_{controls}$
Total	$bb_{cases} + bb_{controls}$	$Bb_{cases} + Bb_{controls}$	$BB_{cases} + BB_{controls}$	N

Table 2.4. Genotype contingency table

The effect of an allele or genotype on a trait is frequently reported as an odds ratio (OR); the odds ratio for an allele or genotype can be estimated by the following:

Allele "b" from Table 2.3 for example,

$$OR_b = (b_{cases} \times B_{controls})/(B_{cases} \times b_{controls})$$

Similarly, the OR for the bb genotype compared to the BB genotype (Table 2.4) is

$$OR_{bb} = (bb_{cases} \times BB_{controls})/(BB_{cases} \times bb_{controls})$$

Classical Twin Modelling

The first reported epidemiological study of multiple births was by James Matthews Duncan. In his book (Duncan, 1865) he discusses amongst other things, the influence of maternal age on the frequency of multiple births. The following is a quote from his book:

"… twinning to be the result of the act of conception taking place on the rupture of a Graafian vesicle fortuitously containing a double ovum or two ova, or on the rupture of two Graafian vesicles accidently matured simultaneously"

Two types of twins are usually distinguished – those that arose from a single fertilised egg are monozygotic MZ twins and are genetically identical, while those that arose from two separately fertilised eggs are DZ twins and share on average half of their genetic material.

The classical twin study design compares phenotypic similarity between MZ and DZ twins. For phenotypes influenced by genetic factors it is expected that MZ twins will show a greater resemblance than DZ twins. Curtis Merriman first described the classical twin study design in 1924 with a study of mental abilities in twins (Rende *et al.*, 1990). Twin studies, like other family based study designs, aim to demonstrate familial clustering of a phenotype, which could be a result of their shared genes and/or familial environment. Where twin studies differ from a family design is that, with care, they allow the estimation of the genetic and shared environmental components that contribute to the observed phenotype.

Assumptions and Limitations

In order for a twin study to be valid, some key assumptions must be met. The most important and perhaps contentious being that to the extent that twins have a shared familial environment, MZ and DZ twins, on average, share their environments to the same degree. This is known as the equal environments assumption (EEA). This can be interpreted as both MZ and DZ twins are exposed to the same (often unknown) environmental factors that may contribute to the phenotypic variance. If MZ interact with their environment in a way that is more similar to the environmental interaction

that DZ twins experience, then any estimate of the genetic effect on the phenotype being studied is likely to be inflated. Conversely, where the opposite is true, i.e. DZ twins systematically experience more similar environmental influences than MZ twins with respect to a specific phenotype, the genetic contribution will be underestimated.

Violation of the EEA can occur at pre- and post-term – in utero monochorionic, monoamniotic MZ twins share the same placenta and amniotic sac. One argument is that the co-localisation of both MZ foetuses confers a more similar intrauterine environment than the majority of DZ twins, which do not share a placenta. Conversely, competition between monochorionic, monoamniotic MZ foetuses for resources may drive intrauterine differences such as those experienced in twin–twin transfusion syndrome (Fisk *et al.*, 2009).

Differences in environmental interaction between MZ and DZ twin pairs may continue from childhood into adulthood, where perceived zygosity may influence behaviour and preferences. MZ and DZ twins may have identical or opposite, for example, dietary habits and physical activity levels as they or their parents (unconsciously) augment or lessen, similarity between them. Another assumption of the classical twin study is that the total observed phenotypic variance must be the same for MZ and DZ twins. Equal variances for MZ and DZ twins indicates the two groups represent a single homogenous group.

Twins are assumed to be representative of the population from which they are sampled from, so that inferences made about the relative contributions to the phenotype can be generalised. In practical terms, this means for continuous phenotypes, such as height or weight, twins and randomly selected individuals from the population should have the same means and variances if the sample size collected is sufficiently large. Similarly, to make claims about a genetic contribution to a disease, the disease prevalence in the twin sample must be the same as in the population from which the twins are drawn from. Twin registries must ensure they do not have a bias in their recruitment such as advertising solely in hospitals, which artificially increases the prevalence of those diseases that require hospital treatment.

Heritability Estimates

Based upon the assumptions outlined above, twin studies allow empirical estimation of the genetic contribution to a phenotype by comparing concordance between MZ and DZ twins. As stated previously, MZ twins are formed by the cleavage of a single fertilised zygote into two separate zygotes, therefore are genetically identical, sharing additive and dominance variances completely. Whilst DZ twins result from the independent fertilisation of two eggs and share, on average, 50% of the alleles at all autosomal loci, making DZ twins no more genetically similar than any other full siblings.

Variance Components

The proportion of dominance variance shared between DZ twins is 0.25, which is the conditional probability that both siblings inherit the same maternal and paternal alleles at a single locus. If subscripts 1 & 2 represent the observed phenotypic value for twin 1 & 2 respectively, the difference in genetic covariance between MZ and DZ pairs can be shown as:

$$Cov_{MZ} = (MZ_1, MZ_2) = V_A + V_D + V_E$$

$$Cov_{DZ} = (DZ_1, DZ_2) = 0.5V_A + 0.25V_D + V_E$$

One way of empirically determining heritability is by structural equation modelling (SEM) that estimates variance components to identify a model that best describes the observed data. These equations take the general form of phenotypic variance decomposition for each twin in a pair:

$$V_{P1} = a_1V_{A1} + d_1V_{D1} + c_1V_{C1} + e_1V_{E1}$$

$$V_{P2} = a_2V_{A2} + d_2V_{D2} + c_2V_{C2} + e_1V_{E2}$$

These equations show the path coefficients (in lower case) for each variance component, the magnitude of these coefficients is not expected to be different between twin 1 and 2 in the pair, thus can be simplified:

$$a_1 = a_2 = a$$

$$d_1 = d_2 = d$$

$$c_1 = c_2 = c$$

$$e_1 = e_2 = e$$

With twin data we want to estimate the relative contributions of the latent genetic and environmental factors, i.e. a variance components model, not the path coefficients *per se*, so all path coefficients are fixed to 1. The variance components are freely estimated with one expected variance-covariance matrix for each model. The variance component parameters that generate the best correspondence between the specified model and the observed variance-covariance matrix represent the best estimate of the true variance component parameters. A SEM equation can be illustrated in the form of a path diagram as presented in Figures 2.7 and 2.8.

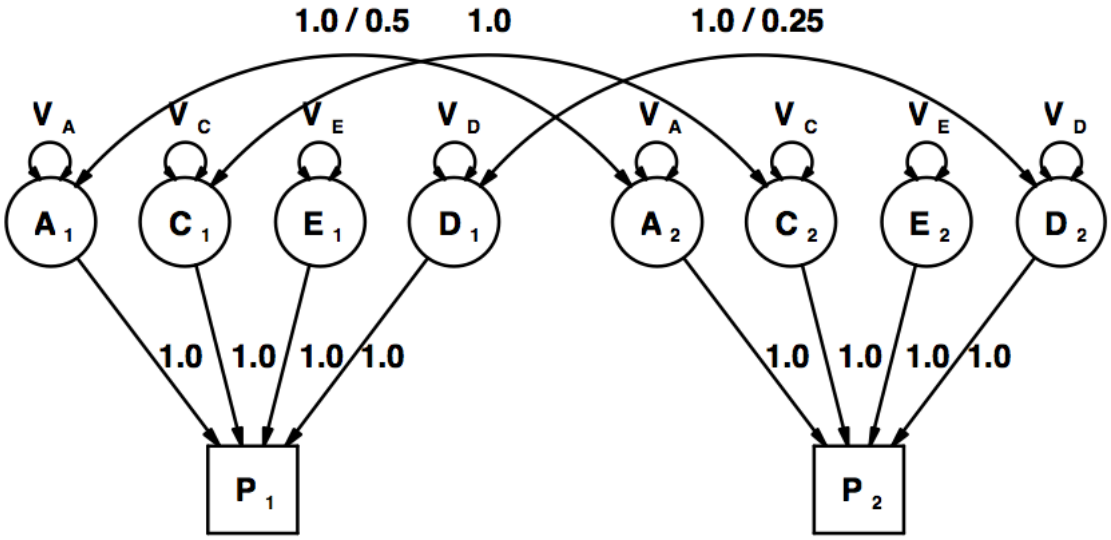


Figure 2.7. Univariate twin path diagram (Maes *et al.*, 1997).
Illustrating the partitioning of phenotypic variance into additive genetic (A), dominant genetic (D), common shared environmental (C), and unique environmental components (E).

In the Figure 2.7 path diagram of twin data, latent genetic and environmental factors with path coefficients point toward the observed phenotype represented by the squared boxes. The unidirectional arrows point causality in that the observed standardised phenotype is caused by unobserved genetic and environmental components. The double-headed arrows represent the covariance of the latent variance components with themselves (thus, the variance of each latent component).

Using tracing rules described by Wright (1934), the total path distance between latent factors and the observed phenotype between twin 1 and 2 is given by:

MZ:

$$P_1 \xleftarrow{1} A_1 \xleftarrow{V_A} \rightarrow A_2 \xrightarrow{1} P_2$$
$$P_1 \xleftarrow{1} D_1 \xleftarrow{V_D} \rightarrow D_2 \xrightarrow{1} P_2$$
$$P_1 \xleftarrow{1} C_1 \xleftarrow{V_C} \rightarrow C_2 \xrightarrow{1} P_2$$

Expected $Cov_{MZ} = V_A + V_D + V_C$

DZ twin 1:

$$P_1 \xleftarrow{1} A_1 \xleftarrow{0.5V_A} A_2 \xrightarrow{1} P_2$$
$$P_1 \xleftarrow{1} D_1 \xleftarrow{0.25V_D} D_2 \xrightarrow{1} P_2$$
$$P_1 \xleftarrow{1} C_1 \xleftarrow{V_C} \rightarrow C_2 \xrightarrow{1} P_2$$

Expected $Cov_{DZ} = 0.5V_A + 0.25V_D + V_C$

The observed covariance matrices are compared to covariance matrices expected from the path diagram model. The expected covariance matrices for MZ and DZ twins are:

For MZ twins:

V_{P1}^{MZ} ($V_{A1} + V_{D1} + V_{C1} + V_{E1}$)	
Cov_{12}^{MZ} ($V_{A12} + V_{D12} + V_{C12}$)	V_{P2}^{MZ} ($V_{A2} + V_{D2} + V_{C2} + V_{E2}$)

For DZ twins:

V_{P1}^{DZ} $(0.5V_A + 0.25V_D + V_{C2} + V_{E1})$	
Cov_{12}^{DZ} $(0.5V_{A12} + 0.25V_{D12} + V_{C12})$	V_{P2}^{DZ} $(0.5V_{A2} + 0.25V_{D2} + V_{C2} + V_{E2})$

For twins that are raised together the effects of genetic dominance (V_D) and a shared environment (V_C) and genetic dominance (V_D) are confounded and cannot be estimated simultaneously (Maes *et al.*, 1997).

When two or more phenotypes are measured for each individual, a multivariate model fitting approach can be performed by extending the univariate model to include the cross-trait, cross twin and cross-trait within individual covariance terms. The path diagram for the model-fitting approach illustrated in Figure 2.8 appears more complicated, but is based upon the same variance component methods.

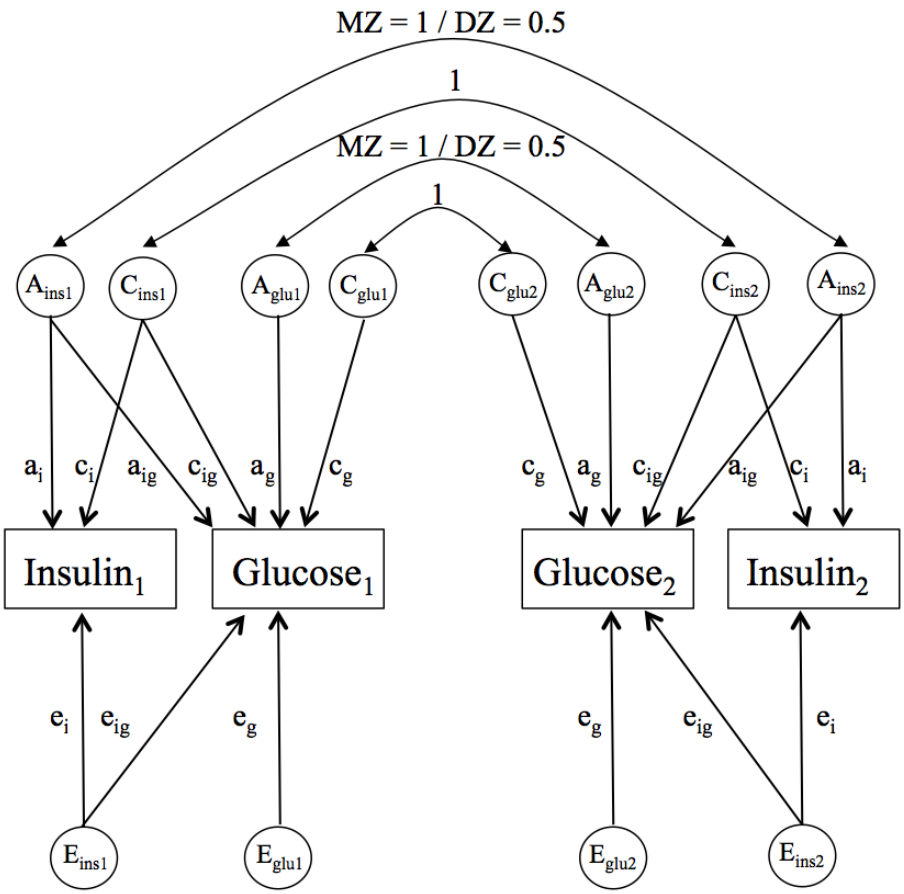


Figure 2.8. A bivariate biometric path diagram.
This pathway diagram shows all latent sources of phenotypic variance (additive genetic [A], common shared environmental [C], and unique environmental [E]) for insulin and glucose values measured for pairs of twins (subscripts 1 and 2). Lower case arrow labels and subscript are factor loadings for each of the variance components.

Linkage Disequilibrium

Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci. For example, consider a pair of biallelic loci A and B with alleles A_1 and A_2 at locus A and B_1 and B_2 at locus B. If the product of the individual allele frequencies of A_1 and B_2 is equal to the observed frequency of the co-occurrence of the alleles at these loci (X_{12}), i.e. $X_{12} - A_1B_2 = 0$, then the alleles are independent and in as state of linkage equilibrium.

Accordingly, when the observed frequency of the co-occurrence of alleles at a pair of biallelic loci is not equal to their expected frequency given by the product of individual allele frequencies, alleles at the two loci are not independent and this disequilibrium is represented by:

$$D_{12} = X_{12} - A_1B_2$$

Where, D_{12} is the difference between the observed frequency co-occurrence of these alleles and the product of allele frequencies at both loci. Changes in D reflect changes in the non-random association between alleles and/or changes in allele frequencies, which is not desirable when comparing different samples. As a solution to this problem, Lewontin (1964) suggested adjusting D by observed allele frequencies to give a relative measure of D, D':

$$D'_{12} = D_{12} / D_{max}$$

Where the dominator of this ratio, D_{max}, can take one of two values depending on the value of D. When D_{12} is less than 0, D_{max} is the minimum value between A_1B_2 and the product of $(1-A_1)(1-B_2)$: $\min[A_1B_2, (1-A_1)(1-B_2)]$.

When D_{12} is greater than 0, D_{max} is the minimum value between $A_1(1-B_2)$ and $B_2(1-A_1)$: $\min[A_1(1-B_2), B_2(1-A_1)]$

Like D, D' remains dependent on allele frequencies but is scaled to allow meaningful comparison between markers when the allele frequencies are not the same. A D' value

of 1 is interpreted as complete LD, while a value of -1 for D' indicates the presence of a very rare allele for the pair of markers being investigated.

D can also be scaled by the observed allele frequencies at both loci (Hill and Robertson, 1968) and the resulting quantity equivalent to the Pearson product-moment correlation coefficient, *r*:

$$r_{12} = D_{12} / \sqrt{(A_1A_2B_1B_2)}$$

The correlation coefficient inherits the sign of D, which can be removed by replacing D with D^2,

$$r^2 = D^2 / \sqrt{(A_1A_2B_1B_2)}$$

The product of the squared correlation coefficient (r^2) and the sample size (n) is equal to Chi-squared (χ^2) test statistic with one degree of freedom: $\chi^2 = r^2n$ for independent loci (Hill and Robertson, 1968). In addition, a χ^2 goodness of fit test could be used to test for equilibrium between markers. The expected allele frequencies (p) assuming independence can be represented by:

		Locus 1		
		A	**A**	
Locus 2	**B**	*pAp*B	*pB*(1- *pA*)	*pB*
	b	*pA*(1- *pB*)	(1- *pA*)(1- *pB*)	1- *pB*
		pA	1- *pA*	

And in the presence of disequilibrium between loci:

		Locus 1		
		A	**A**	
Locus 2	**B**	*pAp*B + D	*pB*(1- *pA*) – D	*pB*
	b	*pA*(1- *pB*) - D	(1- *pA*)(1- *pB*) + D	1- *pB*
		pA	1- *pA*	

The Pearson's χ^2 test is:

$$\chi^2 = \Sigma(O_i - E_i)^2 / E_i^2$$

Where, O_i and E_i are the observed and expected cell counts in the above 2 by 2 contingency tables. The calculated χ^2 value is compared to the critical χ^2 value with 1 degree of freedom for a given level of significance.

Ultimately recombination between loci is the main driver of breakdown in LD (figure 1), the frequency of recombination between loci that are in close proximity is low, consequently it is expected that the greater the distance between loci the greater the frequency of recombination. The recombination fraction (θ) represents the proportion of recombination events between two loci, and so defines genetic distances whereby a 1% recombination frequency between loci equals 1 centiMorgan (cM) on a genetic map, which unlike physical distances, are not additive.

The rate of LD decay over *t* generations depends on the recombination frequency (Figure 2.10) and its magnitude is diminished by a factor of $(1 - \theta)^t$ in each generation (Hedrick, 1987). After *t* generations the LD that remains is:

$$D_t = (1 - \theta)^t D_0$$

Where, D_t is the LD after *t* generations, θ is the recombination fraction, D_0 is the LD in the current generations, and *t* is the number of generations.

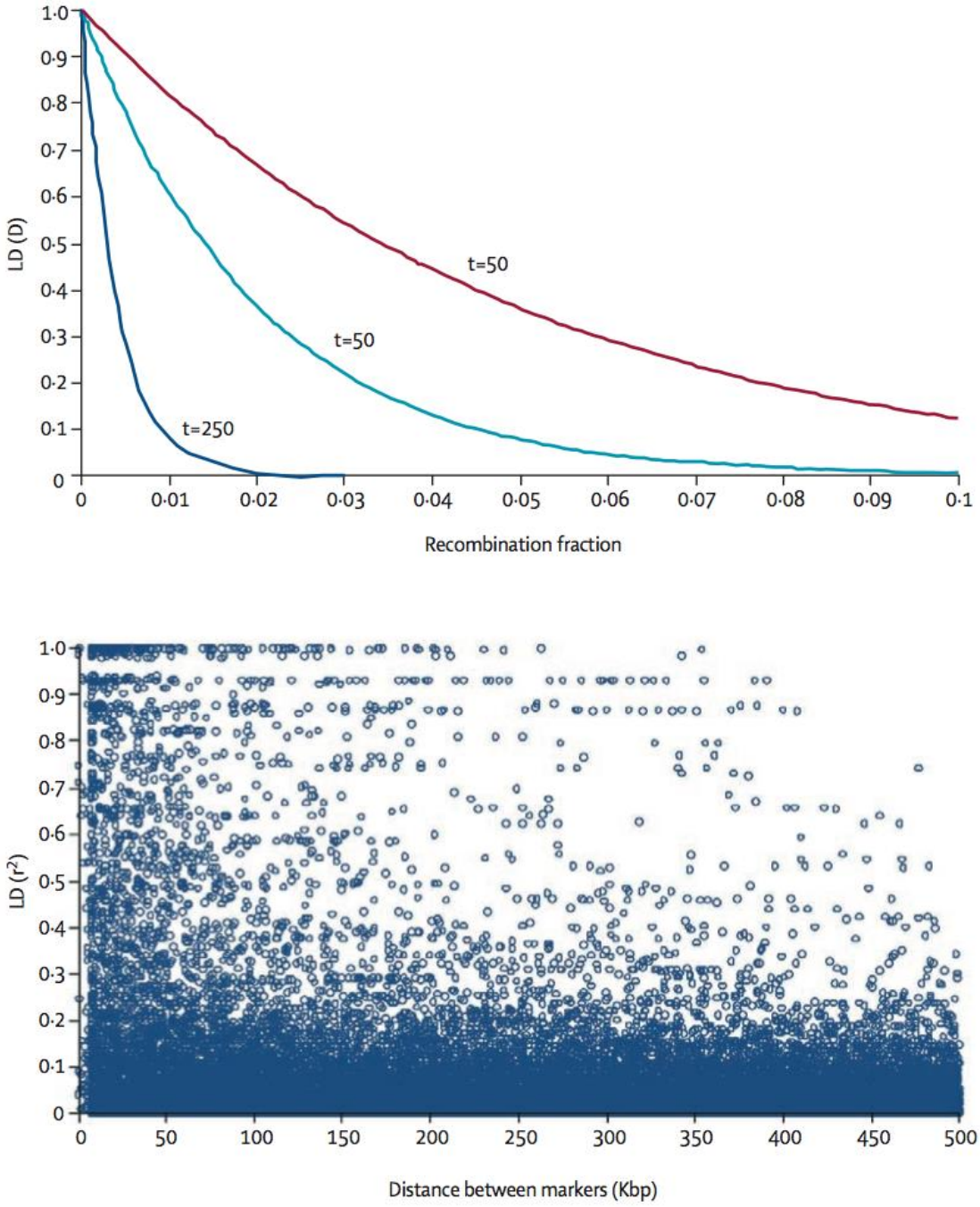


Figure 2.9. Decay of LD with increasing genetic and physical distance (Palmer and Cardon, 2005).

An alternative measure of LD between loci is the association probability (ρ):

$$\rho = D / Q(1-R)$$

Where, D is the difference between observed frequency for a pair of loci and that frequency expected from the product of their individual frequencies. Q and R are the frequencies of the common and rarest alleles for a pair of biallelic loci respectively, which can be illustrated by a 2 by 2 table (Table 2.5).

		SNP 2		**Row Total**	**Frequency**
		1	**2**		
SNP 1	**1**	a	b	a + b	Q
	2	c	d	c + d	1-Q
Column Total		a + c	b + d		
Frequency		R	1-R		

Table 2.5. Frequencies of a pair of biallelic loci (Kuo *et al.*, 2007).
Cells a-d are haplotype frequencies for all combinations for a pair of biallelic SNPs. The table is rearranged so the Q is less than (1-Q, R, 1-R) and the product of haplotype frequency of a+d is greater than b+c.

Collins and Morton (1998) use the Malécot model of isolation by distance (Malécot, 1970) to obtain the expected association ($\hat{\rho}$) between loci:

$$\hat{\rho} = (1 - L)M \exp(-\sum \varepsilon_i d_i)$$

Where,

L reflects residual LD at large distance not due to linkage,

M is the proportion for which the youngest haplotype is monophyletic,

ε is the exponential decline with distance d between a pair of loci.

The parameter, ε, is calculated by considering all pair-wise SNP associations in a region of ordered SNPs. The association between the first and second SNPs is the most informative and so has the largest weighting, the association between the first and third SNPs will have the second largest weighting (Tapper, 2007). The term $\varepsilon_i d_i$ give an additive measure of LD for each SNP, and used for the construction of LD maps (Maniatis *et al.*, 2002). Parameters ε, M, and L are not known but are estimated by maximising the composite likelihood based on the observed association probability (Maniatis *et al.*, 2002).

Chapter 3: The Relationship Between DXA-based and Anthropometric Measures of Visceral Fat and Morbidity in Women.

Abstract

Excess accumulation of visceral fat is a prominent risk factor for cardiovascular and metabolic morbidity. While CT is the gold standard to measure visceral adiposity, this is often not possible for epidemiological studies - thus proxy measures of visceral fat are required. Study aims were to a) identify a valid proxy measure of VAT area, b) estimate VAT heritability and c) assess visceral fat association with morbidity in comparison to body fat distribution. A validation sample of 54 females measured for detailed body fat composition - assessed using CT, DXA, and anthropometry – was used to evaluate previously published predictive models of CT-measured visceral fat. Based upon a validated model, we realised an estimate of abdominal VAT area for a population-based sample of 3,457 female volunteer twins and estimated VAT area heritability using a classical twin study design. Regression and residuals analyses were used to assess the relationship between adiposity and morbidity.

Published models applied to the validation sample explained >80% of the variance in CT-measured visceral fat. Narrow sense VAT area heritability was estimated to be 58% (95% CI: 51-66%) with a shared familial component of 24% (17-30%). VAT area is strongly associated with T2D, hypertension (HT), subclinical atherosclerosis and liver function tests. In particular, VAT area is associated with T2D, HT and alanine aminotransaminase liver function, conditional upon association with DXA total abdominal fat and BMI. DXA and anthropometry measures can be used to derive reliable estimates of visceral fat. Visceral fat is heritable and appears to mediate association between adiposity and morbidity. This observation is consistent with hypotheses that suggest excess visceral adiposity is causally related to cardiovascular and metabolic disease.

Introduction

For the purpose of work presented in the subsequent chapter, it is important to identify the measure of adiposity that bears the strongest relationship with T2D. Mitochondrial dysfunction and their loss in adipocytes through lipid overload may drive excess lipid storage toward insulin resistance (Kusminski and Scherer, 2012), however, the focus of this Chapter is to identify a measure of adiposity to act as an intermediate phenotype for T2D.

Body fat distribution, particularly abdominal adiposity, is strongly associated with a range of chronic metabolic and cardiovascular morbidities. The abdomen includes depots of both subcutaneous and visceral adipose tissue, which are distinct tissues that exhibit different gene expression (Palou *et al.*, 2009), endocrinological profiles (Bastard *et al.*, 2008) and pathogenicity (Kim *et al.*, 2007). The detrimental effect of excess fat accumulation is contingent on body fat distribution. It has been shown that the diabetic phenotype of the leptin knock out obese mice can be rescued by shifting fat storage away from the viscera and into the subcutaneous depot (Kim *et al.*, 2007). The high ratio of subcutaneous to visceral fat in this mouse model is in contrast to the familial partial forms of lipodystrophy observed in humans, a pathology characterised by a redistribution of adipose tissue to the intra-abdominal region, giving rise to severe metabolic abnormalities. These extreme cases illustrate the potent effect that adipose distribution can have upon health (Mantzoros *et al.*, 2011, Savage *et al.*, 2010) and the role of visceral fat accumulation in the aetiology of cardiovascular and metabolic disease.

Adiposity can be measured with a variety of techniques such as DXA, CT, and magnetic resonance imaging (MRI). In whole body composition analysis, CT is regarded as the gold standard and as such has been used extensively as a benchmark for the quantitative assessment of subcutaneous and visceral adiposity. There is a continued effort to identify informative adiposity measures that are easily acquired and show robust correlation with CT-measured visceral fat content (Bertin *et al.*, 2000, De Lucia Rolfe *et al.*, 2011, Hill *et al.*, 2007, Snijder *et al.*, 2002).

Family and twin studies can be used to estimate the heritability (the proportion of trait variance explained by genetic factors) and the relative importance of genetic and shared environmental factors in influencing phenotypic variance, by contrasting trait covariance between relatives with degree of relatedness. Establishing the heritable basis of a trait is also an important consideration prior to gene mapping studies (Andrew *et al.*, 2006).

The aims of this study are to 1) identify and estimate a valid DXA and/or anthropometric based measure of visceral fat, 2) to estimate the heritability of visceral fat accumulation and 3) to assess whether association between body composition measures of adiposity and morbidities can be fully explained by the mediation through visceral fat. We hypothesise that if visceral fat is casually related to morbidity, we would expect visceral fat to remain associated with morbidity independent of other measures of body fat.

Materials and Methods

Subjects and Data Collection

Background information for the subjects used in this study is presented in Chapter 2 (TwinsUK sample). While the twins are largely representative of the age-matched UK female population (Andrew *et al.*, 2001), our findings relating to adiposity are restricted to Caucasian women over 40 years old. The validation sample comprised 54 female individuals over the age of 40 years with both CT and DXA data, but was otherwise unselected. Previously reported linear regression models (Treuth *et al.*, 1995, Snijder *et al.*, 2002, Hill *et al.*, 2007) and adiposity indices (Bertin *et al.*, 2000) were assessed for this sample to identify the most predictive combination of DXA abdominal fat and anthropometry measures required to estimate visceral fat. A validated best-fit model was used to estimate abdominal visceral fat for a study sample of 3,457 individuals (533 MZ twin pairs, 1,102 DZ twin pairs, 187 singletons), where date-matched DXA abdominal fat and anthropometric data were available.

CT

Full details of the CT scans are described in Chapter 2 (TwinsUK sample). The body cavity CSA was obtained by tracing the outer contour of the abdominal wall and adipose tissue within this area was identified as having an attenuation value between -190 and -30 Hounsfield units (Yoshizumi *et al.*, 1999). For the 54 subjects with CT and DXA scan data (referred to as the "validation sample"), visceral fat was directly measured as a pixel count using CT single slice area (VAT area). VAT area was defined as (visceral pixel count / total body cavity pixel count) x CSA. Single slice visceral fat area and volume are highly correlated (Shen *et al.*, 2004, Irlbeck *et al.*, 2010).

DXA

Full details for DXA scan are described in Chapter 2 (TwinsUK sample). A total of 3,457 DXA scans were available for this study (referred to as the "study sample"). The region of interest for body composition analysis was manually defined in a similar manner to Bertin *et al.* (2000); the abdominal region was delineated by an

upper horizontal border located at the lowest rib and a lower border determined by the external end of iliac crests, but the lateral borders were delineated by the inner body wall rather than the body wall. DXA fat mass from this abdominal region was recorded and abdominal transverse internal and external diameters were measured as described by Bertin *et al.* (2000).

Anthropometry

Anthropometric measurements for both the validation and study samples were height, weight, and BMI. Where anthropometric measurements were unavailable for the validation sample they were measured directly from the CT scans and are regarded as anthropometric. Sagittal depth, body cavity, subcutaneous fat and total abdominal fat cross sectional areas, transverse internal and external diameters (TID and TED respectively), subcutaneous fat width (SFW), and waist circumference were measured using CT scans. Skin fold was estimated using the formula SFW = (TED – TID)/2 (Bertin *et al.,* 2000) as calliper skin fold measure was not taken for this study. The validation sample were also measured for TED and TID (with SFW derived) using DXA software to draw a horizontal line on the DXA image at the upper most point of the iliac crest, which was converted from pixel to mm length by multiplying by a constant of 2.048 (DXA software support, Vertec Scientific Limited).

T2D

Subjects were identified as type 2 diabetic if they were classified as hyperglycaemic for one or more of the following diagnostics: fasting glucose serum concentration > 7 mmol/L, fasting oral glucose tolerance test > 11.1 mmol/L or glycosylated haemoglobin > 6.5% (48 mmol/mol). Questionnaire data on physician diagnosis and medication was also used to identify self-reported cases of T2D. Fasting plasma glucose was measured by enzymatic colorimetric slide assay (Johnson and Johnson Clinical Diagnostic Systems, Amersham UK).

Hypertension

Blood pressure was measured twice using an automatic blood pressure monitor (Omron Healthcare, Inc., Bannockburn, IL). Individuals were classified as having

hypertension (HT), if currently receiving anti-hypertensive medication and/or the repeated systolic/diastolic blood pressure was greater than 140/90 mm Hg.

Carotid Intima-Media Thickness

Carotid intima-media thickness (cIMT) was measured as previously described by Cecelja *et al.* (2010) and the quantitative trait was dichotomised at the 75th percentile as a surrogate for subclinical atherosclerosis (Stein *et al.*, 2008, Benzaquen and Nguyen-Thanh, 2009). In brief, the left and right carotid and femoral arteries were visualized with B-mode ultrasound (Siemens CV70, Siemens, Erlangen, Germany, with 13-MHz vascular probe). Common cIMT was measured in the near and far walls, 1 to 2 cm proximal to the carotid bifurcation with automated wall-tracking software (Medical Imaging Applications, Coralville, Iowa) during diastole in an area free of overt plaque. Mean values of cIMT in the near and far walls of both arteries were used for analysis.

Liver Function Tests

Data on liver function test (LFT) proteins were available for a sub-sample of subjects (Rahmioglu *et al.*, 2009). Abnormal liver function as assessed by LFTs can be used to assist diagnosis of non-alcoholic fatty liver disease (Armstrong *et al.*, 2012). Serum concentrations of alanine aminotransaminase (ALT), alkaline phosphatase (ALK), total bilirubin (BIL) and gamma glutamyl transpeptidase (GGT) were used in this study as markers of liver function. Of the four tests, ALT is the most sensitive marker for liver cell damage (as a consequence of disease or drug use), while ALK and GGT are more indicative of cholestatic injury and BIL, haem catabolism (Aragon and Younossi, 2010, Burke, 2002). The upper limit of normal threshold used for each protein were: ALT: 39.4 IU/L (Kariv *et al.*, 2006), ALK: 81 IU/L (taken as the 75% percentile of the sample distribution), BIL: 17.1 □mol/L (Lazo *et al.*, 2008), GGT: 33 IU/L (Sabanayagam *et al.*, 2009).

Model Validation

Multiple regression analyses (Treuth *et al.*, 1995, Snijder *et al.*, 2002, Hill *et al.*, 2007) and indices of adiposity were assessed (Bertin *et al.*, 2000) to address two questions relating to the estimation of total abdominal visceral fat using DXA

adiposity and a range of anthropometric measures: 1) which of these predictive models for the "gold standard" CT measure of visceral fat, best fit our validation sample of 54 females; and 2) in relation to previous discussions (Bertin *et al.*, 2000), can visceral fat be reliably estimated based upon anthropometry alone. CT and DXA scans for the same individuals were date-matched to between 0.23 - 2.5 years of one another. The difference in scan date for the validation sample was included in all visceral fat regression models as a nuisance factor. A Bland-Altman analysis was conducted to assess if the predicted VAT error term was constant or varied across the range of CT-measured VAT area.

Heritability Analysis

The classical twin model was used to estimate the relative influence of genetic and environmental factors upon individual variation about the visceral fat sample mean (Martin *et al.*, 1997). Using variance components analysis, the total phenotypic variance was partitioned into estimates of the additive (A), dominant (D) genetic and shared familial (common; C) and unique to the individual (E) environmental (and measurement error) components using genetically informative data. The model assumes no epistasis, gene-environment correlation or interactions and that shared environmental factors are not confounded by zygosity (the equal environments assumption). Provided these assumptions hold, on the basis that MZ twins share identical genes and DZ twins share, on average, half their segregating genes, twin data can be used to infer heritability and shared familial estimates. Mx (Neale *et al.*, 1992) was used for all heritability analyses.

In addition to the univariate heritability analysis, we also conducted a bivariate variance components analysis between the realised estimate of VAT area and total abdominal fat as an indirect means to further validate the proxy measure of VAT area. Bivariate analysis facilitates a test of genetic correlation between two variables, but also whether each variable has a specific genetic component that is not shared by the two traits. We used the specific test to assess if there was evidence for a genetic component that is unique to VAT area and not shared with DXA total abdominal fat (Neale *et al.*, 1992).

Morbidity Association with Adiposity

T2D, hypertension and liver function were associated with the different measures of adiposity, age and other confounding variables using logistic regression. Incident subclinical atherosclerosis (cIMT) was modelled using Cox regression and baseline measures of visceral fat taken on average 10 years (range 5 - 16 years) previously. Since there was no baseline examination of cIMT, we could not exclude the small proportion of individuals who may have already developed pre-clinical atherosclerosis at baseline. To the extent this was true, the study design for these individuals would have been cross-sectional rather prospective.

LFTs were performed using blood samples collected at the same visit as the DXA scan. The mean, standard deviation and median were all examined to identify potential batch effects across years, but no obvious trend was identified. Therefore the year of visit was categorised as quintiles and included in all analyses as a categorical confounding variable.

Co-linearity between the explanatory measures of VAT area, DXA total abdominal fat, BMI and age was assessed using pair-wise correlations and residuals analysis (David Clayton, 1993, O'Brien, 2007). A variance inflation factor (VIF) was calculated for these data, with a score of > 10 sometimes used as a (conservative) threshold indicator for potentially problematic co-linearity between model explanatory variables (O'Brien, 2007). To assess consistency of results, the multiple regression analyses were repeated using random subgroups of data and using quantitative traits for all subjects, where morbidity had been defined using a quantitative trait threshold (i.e. systolic and diastolic blood pressure and LFTs).

Parsimonious best-fit multiple regression models for morbidity were identified by a) using a likelihood ratio test (LRT) to assess the contribution of each explanatory variable to the full model; b) assessing model fit using pseudo-R^2 for T2D, HT and LFTs and the Wald statistic for incident cIMT, and c) the implementation of a linear residuals analyses (David Clayton, 1993) to assess whether adiposity measures are independent of morbidity when conditioned upon on visceral fat. For example, to assess if VAT completely mediates the association between BMI and T2D, secondary residuals analysis involved taking the residuals for T2D on the logit scale and the

ordinary least squares residuals for BMI, both with respect to VAT area and age. Ordinary least squares (OLS) residuals for DXA total abdominal fat were also taken with respect to VAT area and age. The secondary residuals analysis then involved the linear OLS regression of T2D upon BMI residuals and DXA residuals. No evidence of association between residuals would imply that BMI/DXA association with morbidity is primarily mediated via VAT area, while a significant residual association demonstrates either that BMI/DXA are associated with morbidity through a path not mediated by VAT area, or that the parametric assumptions of the association models are false.

Unadjusted and adjusted OR results are presented in the results for VAT area (cm^2), DXA total abdominal fat (kg) and BMI (kg/m^2). To facilitate comparison between the different adiposity units, all adipositities were standardised for the regression analyses. Robust standard errors were estimated by grouping twin pairs using the cluster option in Stata to account for intra-family relatedness. All statistical analyses were performed using Stata version 11.1 (StatCorp, Texas). The VIF and residuals analyses were performed by Dr Toby Andrew, full analyses are presented for T2D (Supplementary Tables 3.2 and 3.3) as an example but were performed for all morbidities.

Results

Estimation of Visceral Fat (Validation Sample)

Descriptive statistics for anthropometric and body composition for the validation and study samples are presented in Table 3.1. All subjects were female and over 40 years old at examination. The mean height, weight, BMI and DXA abdominal fat do not statistically differ between the two samples, although the mean age for the validation sample was six years older (p-value = 6.5×10^{-11}) with a larger waist circumference (p-value = 2.05×10^{-6}) than the study sample. The study sample prevalence and 95% confidence interval for morbidity related to the quantitative traits presented in Table 3.1 were as follows: T2D 0.046 (0.04-0.05), HT 0.08 (0.07-0.09), cIMT 0.27 (0.24-0.30), ALT 0.22 (0.21-0.24), ALK 0.27 (0.26-0.29), BIL 0.02 (0.02-0.03) and GGT 0.20 (0.18-0.21).

	Validation sample (n = 54)				Study sample (n = 3457)			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Age (years)	60.4	6.1	49.3	72.8	54.2	8.3	40.0	79.5
Weight (kg)	65.7	9.4	48.4	87.4	66.6	11.8	35.6	139.5
Height (m)	1.62	0.06	1.48	1.75	1.62	0.06	1.39	1.82
Waist circumference (cm)	88.0	9.9	66.6	111.2	81.0	11.0	55.0	134.0
Sagittal depth (cm)	21.8	3.2	15.9	31.1	-	-	-	-
Scan difference (years)	1.3	0.8	0.2	2.5	-	-	-	-
BMI (kg/m^2)	25.1	3.8	19.2	33.8	25.6	4.5	15.1	51.7
Total abdominal fat (kg)	1.4	0.62	0.24	3.08	1.44	0.61	0.14	3.94
VAT area (cm^2)	127.8	52.1	37.7	279.5	144.6	49.6	37.7	347.4
Diastolic BP (mmHG)	77.3	8.3	61.0	95.5	75.9	8.9	47.5	108.0
Systolic BP (mmHG)	122.0	11.5	92.0	151.0	123.1	14.7	86.5	189.0
cIMT	0.68	0.08	0.53	0.82	0.67	0.11	0.30	1.11
ALT	23.9	11.8	3.0	68.0	26.7	11.4	2.5	217.3
ALK	64.8	19.0	26.0	114.0	71.2	18.2	23.5	218.9
BIL	9.5	3.7	5.7	23.5	8.7	3.0	1.0	30.5
GGT	30.3	17.4	12.0	65.3	27.9	21.7	3.0	359.0

Table 3.1. Validation and study sample characteristics.
Age, weight, height, body fat distribution and intermediate quantitative traits used to define clinical morbidity. Waist circumference for the validation sample is based upon the transverse circumference of CT body scan image at the waist, while the study sample is a tape measurement taken at the same time as the DXA scan. Abbreviations: ALT – alanine aminotransaminase, ALK - alkaline phosphatase, BIL – bilirubin, BP – blood pressure, cIMT – carotid intima-media thickness, GGT - gamma glutamyl transpeptidase, SD – standard deviation.

The Pearson product moment correlation coefficients (*r*) between the different adiposity measures and CT-measured VAT area for the validation sample are presented in Table 3.2. CT-measured VAT area was most strongly correlated with CT-measured body cavity cross sectional area ($r = 0.85$), sagittal depth ($r = 0.84$) and tape-measured waist circumference ($r = 0.86$) and DXA total abdominal fat ($r = 0.79$). Consistent with these data, is the observation that reported models of visceral fat in the literature, whether linear regressions or derived anthropometric indices, all attempt to capture information about the body cavity volume (or area) in relation to the subcutaneous volume (Treuth *et al.*, 1995, Bertin *et al.*, 2000, Snijder *et al.*, 2002, Hill *et al.*, 2007). This insight was used to guide the choice of linear regression to estimate CT-measured visceral fat.

	VAT area	BC	Sub. CSA	Total CSA	SD	WC	TID	TED	SFW	DXA	BMI
BC	**0.85**										
Sub. CSA	0.58	0.44									
Total CSA	0.81	0.80	0.90								
SD	**0.84**	0.80	0.84	0.96							
WC	**0.86**	0.77	0.83	0.94	0.94						
TID	0.66	0.72	0.46	0.67	0.61	0.65					
TED	0.66	0.56	0.89	0.87	0.82	0.85	0.62				
SFW	0.43	0.26	0.83	0.69	0.66	0.67	0.17	0.88			
DXA	**0.79**	0.56	0.68	0.74	0.76	0.77	0.56	0.75	0.60		
BMI	0.71	0.60	0.80	0.84	0.83	0.82	0.57	0.85	0.72	0.79	
Weight	0.67	0.64	0.77	0.84	0.76	0.81	0.69	0.88	0.68	0.69	0.86

Table 3.2. Validation sample (n = 54) correlation coefficients between computed tomography visceral adipose tissue (VAT) area, anthropometric and abdominal fat measures.

All measures presented here are based upon CT scan images apart from DXA total abdominal fat, BMI and weight. Abbreviations and units: BC - body cavity area (cm^2), BMI - body mass index (kg/m^2), CSA – body cross-sectional area (cm^2) at L4:L5, DXA – DXA-measured total abdominal fat (kg), SD - sagittal depth (cm), Sub.CSA – subcutaneous cross sectional area, SFW - subcutaneous fat width (cm), TED - transverse external diameter (cm), TID - transverse internal diameter (cm), VAT area - visceral adipose tissue area (cm^2), WC – waist circumference (cm) derived from the CT-image, weight (kg).

Table 3.3 presents the results for three previously reported DXA-based regression models and anthropometric indices for estimating visceral fat, applied to the TwinsUK CT validation sample. The best-replicated regression models included DXA trunk fat and sagittal depth ($R^2 \approx 0.8$), while a combination of DXA and skin fold was less predictive of visceral fat. The best individual indices were functions of sagittal depth (SD), SFW, TID and TED ($r > 0.85$ or $r^2 > 0.72$). Moreover, the most reproducible indices of visceral fat all relate to body cavity CSA. By assuming body cavity CSA takes the form of an ellipse, this was calculated as BC CSA = π x (SD – 2SFW) x TID for the validation sample.

A		Model	Reported R^2	TwinsUK R^2
	Snijder *et al.* (2002)	DXA trunk fat + sagittal depth	0.74	0.80
		DXA trunk fat + abdominal circumference	0.71	0.78
	Treuth *et al.* (1995)	Sagittal depth + age + waist circumference + % DXA trunk fat	0.81	0.79
	Hill *et al.* (2007)	DXA + skin fold	0.68	0.65
B		**Index**	**Reported r**	**TwinsUK r**
	Bertin *et al.* (2000)	Abdominal fat mass (kg)	0.57	0.79
		Thigh fat mas (kg)	0.06*	-
		Abdominal fat mass / thigh fat mass	0.75	-
		Abdominal fat mass / SFW	0.83	0.58
		TED (cm)	0.54	0.61
		TID (cm)	0.9	0.61
		SFW (cm)	-0.23*	0.28
		(SD)(TID)	0.89	0.87
		(SD)(TID) / height	0.91	0.86
		(SD)(TID) / BMI	0.66	0.49
		(SD-SFW)	0.86	0.89
		(SD-SFW)(TID)	0.92	0.79
		(SD-SFW)(TID) / height	0.94	0.87

Table 3.3. Visceral adipose tissue (VAT) area linear model estimates and correlational indices.

Previously reported models in the literature were applied to the TwinsUK validation sample of CT-measured VAT area (n=54) and the coefficient of determination (R^2) is presented for each study as an indication of the proportion of VAT variance explained by the model (**A**). The Pearson product-moment correlation coefficients between CT-measured VAT area and adiposity indices described in Bertin *et al.* (2000) were also calculated for the TwinsUK validation sample (**B**). Asterisks in B indicate the reported correlation coefficient does not differ significantly from zero (at significance threshold = 0.05). Abbreviations: DXA - dual-energy X-ray absorptiometry, SD - sagittal depth, SFW - subcutaneous fat width, TED - transverse external diameter, TID - transverse internal diameter.

In modelling CT-measured VAT area, the best fit and most interpretable model included a combination of measures for DXA abdominal fat, body cavity cross sectional area (estimated using the ellipse formula above) and waist circumference (Table 3.4, Model 0), which together explained 91% of the variance in CT-measured VAT area (R^2 = 0.91). However, since sagittal depth was not available for the study sample for which we wished to estimate VAT area, we also assessed a model including only DXA total abdominal fat, tape-measured waist circumference and age. This model obtained an R^2 of 0.83 (Table 3.4, Model 1) and is described by the following linear regression equation using standardised explanatory variables and no intercept term:

VAT area = 10.1(DXA abdominal fat mass, kg) + 40.8(waist circumference, cm) + 1.4(age, years)

For this estimate, a Bland-Altman analysis showed no evidence of heteroscedascity across the full range of CT-measured VAT area, with only 2 out of 54 (3.7%) values outside the 95% limits of agreement (the mean difference ± twice the standard deviation of the difference between the two measures). Model 1 was selected for heritability and morbidity analyses.

VAT Model	Measure	β	SE	*t*	p-value	Lower 95% CI	Upper 95% CI	Model R^2
(A) Model 0								0.91
Combination of DXA & anthropometric measures	DXA abdominal fat	20.07	3.39	5.9	$2x10^{-9}$	13.2	27.0	
	BC CSA	32.44	4.52	7.2	$4x10^{-13}$	23.2	41.6	
	WC	11.07	5.57	2.0	$2x10^{-2}$	-0.3	22.4	
(B) Model 1								0.83
Combination of DXA & anthropometric measures	DXA abdominal fat	10.1	4.82	2.1	0.04	0.31	19.9	
	WC	40.8	5.68	7.2	$3x10^{-13}$	29.2	52.3	
	Age	1.4	0.52	2.6	0.01	0.32	2.44	
(C) Model 2								0.86
Anthropometric measures only	BC CSA	25.5	5.58	4.57	$2x10^{-6}$	14.1	36.8	
	WC	30.5	5.46	5.59	$1x10^{-8}$	19.4	41.6	

Table 3.4. Linear regression models for computed tomography (CT) visceral adipose tissue (VAT) area using the validation sample (n = 54).

(A) Model 0: combination of DXA and anthropometric measures guided by previously published models presented in Table 3.3A. **(B) Model 1**: combination of DXA and anthropometric measures restricted to DXA total abdominal fat, WC and age that were also available for the study sample. **(C) Model 2**: using anthropometric measures only. BC CSA was estimated using BC = (π x [SD–2SFW] x TID) from the CT images at intervertebral disc L4:L5 as described in Chapter 2 (TwinsUK sample). Note that for Model 2, using these explanatory variables instead of BC CSA, yields equally good or better prediction of VAT area (R^2 = 0.89), but the model is less interpretable with a negative beta coefficient for SFW. Abbreviations: BC - body cavity, CSA - cross sectional area, DXA - dual-energy X-ray absorptiometry, SD - sagittal depth, SFW - subcutaneous fat width, TED - transverse eternal diameter, TID - transverse internal diameter.

In relation to efforts attempting to estimate visceral fat using only anthropometric measures (Bertin *et al.*, 2000, Karelis *et al.*, 2012), we also obtained a highly explanatory model (R^2 = 0.86) with a linear equation using only two CT measures of body cavity CSA (estimated as an ellipse) and waist circumference (Table 3.4, Model 2). This figure rose to R^2 = 0.89 using body cavity components SD, SFW and WC as explanatory variables for CT VAT area (data not shown). Body cavity CSA results are presented, as this model is more interpretable, while the model including SD, SFW and WC yields a negative beta coefficient for SFW due to co-linearity. Again, these simple anthropometric models could not be used for the study sample because SD was not recorded.

In addition to these validation models, we indirectly assessed the validity of our study sample estimates of visceral fat by making two observations:

1. Realised estimates of VAT area (VAT_{Model1} and VAT_{Model2}) for the validation sample were equally or more strongly correlated with VAT_{CT} (r = 0.89 and r = 0.93) than DXA total abdominal fat (r = 0.88 and r = 0.70);

2. Bivariate variance component analysis between VAT_{Model1} and DXA total abdominal fat provided strong statistical evidence ($\Delta\chi^2_1$= 43.7, p = $4x10^{-11}$) for a specific heritable component that was unique to VAT_{Model1} and not shared with DXA total abdominal (data not shown).

Heritability

The data provided a narrow sense heritability estimate (A) of 0.58 (95% CI: 0.51-0.66) and a familial environmental effect of 0.24 (95% CI: 0.17-0.30) for VAT area (Table 3.5), with the best fit to the data being the ACE model, including additive genetic (A), common familial (C) and unique environmental (E) components.

Model	-2 LL	df	AIC	Δdf	$\Delta\chi^2$	*p* - value	Component	h^2	95% CI
ACE	51399.9	4737	41925.9	-	-	-	A	0.58	(0.51-0.66)
							C	0.24	(0.17-0.30)
							E	0.18	(0.16-0.20)
AE	51438.7	4738	41962.7	1	39	$4.7x10^{-10}$	A	0.83	(0.81-0.84)
							C	-	-
							E	0.18	(0.16-0.19)
CE	51608.4	4738	42132.4	1	209	$2.9x10^{-47}$	A	-	-
							C	0.62	(0.59-0.64)
							E	0.38	(0.36-0.41)

Table 3.5. Visceral adipose tissue (VAT) area estimate of heritability (h^2) and model fit statistics (n = 3,457).

Full (ACE) and nested (AE and CE) model estimates are presented. Nested models test the hypothesis that the estimated additive polygenic genetic variance component (model CE) and the shared familial environmental component (model AE) do not contribute to the observed phenotypic variance. The full ACE model is highlighted as best-fit, since the more parsimonious nested models do not fit the data as well ($p \leq 0.05$). The model with the lowest AIC fit statistic also indicates best model fit. Abbreviations: -2 LL: minus twice the log-likelihood; AIC: Akaike's Information Criterion; $\Delta\chi^2$: likelihood ratio χ square statistic; A – additive polygenic variance component, C – common familial environment, E – unique environmental variance (and measurement error) specific to the individual.

Visceral Fat as a Risk Factor of Morbidity (Study Sample)

The correlation between study sample explanatory variables is presented in Supplementary Table 3.1. As a function of these variables, VAT area is most strongly correlated with waist circumference and DXA total abdominal fat. While the univariate odd ratios for VAT area, DXA abdominal fat and BMI were all significantly associated with each morbidity (Tables 3.6-3.9), VAT area was most consistently and strongly associated with four morbidity traits of T2D, HT, cIMT and ALT.

For T2D, while the univariate ORs for the 3 adiposity measures were all associated (Table 3.6A), visceral fat and age provided the best-fit multiple regression model (pseudo-R^2 = 0.08, Table 3.6B and LRT, Supplementary Table 3.2), with an adjusted OR of 2.08 (95% CI 1.76 – 2.47) per standard deviation increment in VAT area including age. Removing VAT area from the full model resulted in a significant decline in model fit (LRT χ^2_1 = 14.4, p-value = $1x10^{-4}$), while the removal of DXA total abdominal fat and BMI, either individually or together (χ^2_2 = 1.9, p-value = 0.39) did not reduce the model fit (Supplementary Table 3.2). Residuals analysis, to assess model co-linearity, indicated VAT area mediated the association between T2D and adiposity (Supplementary Table 3.3). No residual association between T2D and DXA and BMI were observed when the residuals for VAT area and age were used (p-value = 0.38), while strong evidence for association between T2D and VAT area remained, when residuals for DXA, BMI and age were used (p-value = $1x10^{-4}$).

Type 2 Diabetes		OR	SE	z	p-value	Lower 95% CI	Upper 95% CI	Model Pseudo-R^2
A	VAT area	2.17	0.18	9.5	$<2x10^{-16}$	1.85	2.54	0.07
	DXA abdominal fat	1.86	0.13	8.6	$<2x10^{-16}$	1.61	2.14	0.05
	BMI	1.66	0.12	7.2	$2x10^{-13}$	1.45	1.91	0.04
	Age	1.05	0.01	4.3	$8x10^{-6}$	1.03	1.07	0.02
B	VAT area	2.08	0.18	8.5	$<2x10^{-16}$	1.76	2.47	0.08
	Age	1.02	0.01	2.0	0.05	1.00	1.05	

Table 3.6. Type 2 diabetes and adiposity.
The study sample prevalence (females > 40 years) estimate for T2D of 5%. **(A)** Logistic regression (n = 2,964) presenting unadjusted OR for each adiposity measure and (**B**) best-fit multiple regression model. For evidence of the presented best-fit model and an analysis of residuals to account for co-linearity between adiposity variables, see Supplementary Tables 3.1 and 3.2, respectively. Explanatory variables VAT area, DXA and BMI are all standardised, implying a change in odds ratio per unit SD change. Abbreviations: BMI - body mass index, CI - confidence interval, DXA - dual-energy X-ray absorptiometry, OR - odds ratio, SE - standard error, VAT - visceral adipose tissue.

Hypertension was equally strongly associated with VAT area, DXA abdominal fat and BMI for univariate analyses (Table 3.7A), but visceral fat and age provided the best-fit multiple regression model (pseudo-R^2 = 0.10, Table 3.7B), with an adjusted OR of 1.90 (95% CI 1.60 – 2.25) per standard deviation increment in VAT area including age. Removing VAT area from the full model resulted in a nominal decline in model fit (LRT χ^2_1 = 7.1, p-value = 0.01), whilst the removal of DXA total abdominal fat and BMI, either individually or together (χ^2_2 = 1.99, p-value = 0.37) did not reduce the model fit. Residuals analysis indicated VAT area mediated the association with HT, with no residual association between HT and DXA and BMI observed when the residuals for VAT area and age were used (p-value = 0.23), while strong evidence for association between T2D and VAT area remained, when residuals for DXA, BMI and age were used (p-value = 0.004).

Hypertension		OR	SE	z	p-value	Lower 95% CI	Upper 95% CI	Model Pseudo-R^2
A	VAT area	2.08	0.16	9.5	$<2x10^{-16}$	1.79	2.42	0.08
	DXA abdominal fat	1.77	0.14	7.4	$6 x10^{-14}$	1.53	2.07	0.05
	BMI	1.77	0.13	7.7	$6 x10^{-15}$	1.53	2.05	0.06
	Age	1.07	0.01	6.4	$6 x10^{-11}$	1.05	1.09	0.05
B	VAT area	1.90	0.17	7.4	$9 x10^{-14}$	1.60	2.25	0.10
	Age	1.04	0.01	4.0	$4 x10^{-5}$	1.02	1.07	

Table 3.7. Hypertension and adiposity.
The study sample prevalence estimate (females > 40 years) for hypertension of 8%. **(A)** Logistic regressions (n = 2,032) showing unadjusted OR for each adiposity measure and **(B)** best fit multiple regression model. For evidence of the presented best-fit model and an analysis of residuals to account for co-linearity between adiposity variables, see Supplementary Tables 3.3 and 3.4, respectively. Explanatory variables VAT area, DXA and BMI are all standardised, implying a change in odds ratio per unit SD change. Year of visit was categorised as quintiles and included in all HT analyses as a categorical confounding variable. Abbreviations: BMI - body mass index, CI - confidence interval, DXA - dual-energy X-ray absorptiometry, OR - odds ratio, SE - standard error, VAT - visceral adipose tissue.

The prospective analysis of subclinical atherosclerosis had a median follow up time of 9.7 years, during which a total of 221 (27%) individuals were classified as sub-clinically atherosclerotic. Univariate Cox proportional hazard models showed all three measures of adiposity and age at baseline to be associated with incident cIMT (Table 3.8A), with VAT area (χ^2_1 = 43.8) and age (χ^2_1 = 40) providing the best-fit parsimonious model (Table 3.8B). DXA and BMI could be dropped from the full model with no nominal (χ^2_2 = 5.7, p-value = 0.06) deterioration in model fit. Residuals analysis indicated VAT area mediated the association with cIMT, with no residual association between cIMT and DXA and BMI observed when the residuals for VAT area and age were used (p-value = 0.19), while strong evidence for association between cIMT and VAT area remains, when residuals for DXA and BMI were used (p-value < $5x10^{-4}$). Both VAT area and BMI better predicted cIMT than DXA total abdominal fat.

Carotid Intima-Media Thickness		HR	SE	z	p-value	Lower 95% CI	Upper 95% CI	Model Fit (Wald)		
								χ^2	df	p-value
A	VAT area	1.50	0.09	6.6	$2x10^{-11}$	1.33	1.69	43.8	1	$4x10^{-11}$
	DXA abdominal fat	1.29	0.07	4.5	$4x10^{-6}$	1.16	1.45	20.0	1	$8x10^{-6}$
	BMI	1.39	0.08	5.5	$2x10^{-8}$	1.23	1.55	30.5	1	$3x10^{-8}$
	Age	1.08	0.01	6.3	$1x10^{-10}$	1.05	1.10	40.0	1	$3x10^{-10}$
B	VAT area	1.36	0.095	4.4	$5x10^{-6}$	1.19	1.56	59.6	2	$1x10^{-13}$
	Age	1.06	0.013	5.0	$3x10^{-7}$	1.04	1.09			

Table 3.8. Sub-clinical atherosclerosis and adiposity.

(A) Cox proportional hazards regression (n = 801) showing unadjusted HR for each adiposity measure and **(B)** best fit multiple regression model. The study sample prevalence estimate (females > 40 years) for sub-clinical atherosclerosis at follow-up was 0.27 (average time from baseline to follow-up was 9.95 years, range 5 - 16 years). Explanatory variables VAT area, DXA and BMI are all standardised, implying a change in hazard ratio per unit standard deviation change. For Cox proportional hazards, the Wald model-fit statistic is presented to indicate the best model fit (StataCorp, Texas) that predicts onset of sub-clinical atherosclerosis (carotid intima-media thickness, cIMT). Abbreviations: BMI - body mass index, CI - confidence interval, DXA - dual-energy X-ray absorptiometry, HR - hazard ratio, SE - standard error, VAT - visceral adipose tissue.

All LFT protein serum levels were positively associated with measures of adiposity (Table 3.9), except for bilirubin, which was negatively associated. VAT area remained associated with alanine aminotransferase (ALT) when conditioned upon DXA total abdominal fat and BMI, while VAT area and DXA were still associated with ALK, BIL and GGT conditional upon BMI (Table 3.9). Removing VAT area from the full model for ALT resulted in a significant decline in model fit (LRT χ^2_1 = 19.6, p-value = $1x10^{-5}$), while the removal of DXA total abdominal fat and BMI, either individually or together (χ^2_2 = 1.56, p-value = 0.46) did not. Residuals analysis indicated that VAT linearly mediates the associations between other adiposity traits and ALT, with no residual association between ALT and DXA and BMI observed when the residuals for VAT area and age were used (p-value = 0.90). Strong evidence for association between ALT and VAT area remained, when residuals for DXA, BMI and age were used (p-value = $1x10^{-4}$).

The estimated variance inflation factor between VAT area and DXA abdominal fat, BMI and age was 8.84. Analyses repeated using subgroups of data and analyses using underlying quantitative traits for HT, ALT, ALK, BIL and GGT all provided qualitatively the same association results (data not shown).

Liver Function Test	Best Fit Model	OR	SE	z	p-value	Lower 95% CI	Upper 95% CI	Model Pseudo-R^2
ALT (0.22)	VAT area	1.75	0.09	10.9	$<2x10^{-16}$	1.58	1.93	0.09
	Age	1.02	0.01	2.9	0.004	1.01	1.03	
ALK (0.27)	VAT area	1.28	0.14	2.4	0.02	1.05	1.58	0.09
	DXA abdominal fat	1.26	0.13	2.3	0.02	1.03	1.54	
	Age	1.06	0.01	9.8	$<2x10^{-16}$	1.05	1.07	
BIL (0.02)	VAT area	0.62	0.10	-3.0	0.003	0.46	0.85	0.054
BIL (0.02)	DXA abdominal fat	0.67	0.10	-2.7	0.01	0.50	0.90	0.049
GGT (0.20)	VAT area	1.25	0.14	1.9	0.05	1.00	1.56	0.06
	DXA abdominal fat	1.36	0.15	2.8	0.01	1.10	1.70	
	Age	1.02	0.01	2.7	0.007	1.00	1.03	

Table 3.9. Liver function tests and adiposity.

Best-fit multiple regression models are presented for the logistic regression models (n = 3,014) including potential explanatory variables VAT area, DXA total abdominal fat, BMI and age. Prevalence for upper limit of normal threshold for each assay is indicated in brackets (Chapter 2: TwinsUK sample). Explanatory variables VAT area, DXA and BMI are all standardised, implying a change in odds ratio per unit standard deviation change. Year of visit was categorised as quintiles and included in all LFT analyses as a categorical confounding variable. Note that two multiple regression models are presented for BIL, since both DXA and VAT area predict BIL equally well. Including both measures in this model provides uninterpretable ORs due to co-linearity between the variables (see Supplementary Table 3.1). Abbreviations: ALT – alanine aminotransaminase, ALK - alkaline phosphatase, BIL – bilirubin, BMI - body mass index, CI - confidence interval, DXA - dual-energy X-ray absorptiometry, GGT - gamma glutamyl transpeptidase, OR - odds raito, SE - standard error, VAT - visceral adipose tissue.

Discussion

This study provides evidence that a validated DXA-based measure of visceral fat is a strong risk factor for common metabolic and cardiovascular diseases. Validation sample explanatory models were able to explain >80% of the variance in CT-measured visceral fat. A combination of DXA and anthropometric measures (R^2 = 0.91) or including only anthropometric measures (R^2 = 0.86), both provided equally good estimates of visceral fat. We obtain a heritability of 58% for visceral fat, which is consistent with familial studies that report within the region of 48-57% (Perusse *et al.*, 1996, Rice *et al.*, 1997, Hong *et al.*, 1998, Fox *et al.*, 2007) and using a twin design, confirms for the first time that the observed familial component is also partly due to shared familial environment (24%).

Our results confirm that visceral fat is the single most important measure of adiposity for risk of T2D. In large DXA resources, such as the one used by Leslie (2010), it would be interesting to ascertain whether the risk of developing diabetes is better predicted if a DXA-based estimate of visceral fat was used rather than using DXA total abdominal fat.

The prevalence of hypertension in the study sample is 8%, which is consistent with the range reported for this demographic by the Health Survey for England (2008). Visceral fat mediates all associations between hypertension and adiposity variables for these cross sectional data. Similarly, in a longitudinal study, Hayashi *et al.* (2004) showed visceral is significantly associated with hypertension and when visceral fat is adjusted for other adiposity measures (e.g. BMI, waist circumference or abdominal subcutaneous fat). Visceral and subcutaneous adiposity both contribute to the prevalence of hypertension, but when adjusted for BMI or waist circumference, subcutaneous fat is no longer associated with hypertension (Fox *et al.*, 2007). This further illustrates that it is the distribution between these two fat depots that determines the risk of morbidity.

The LFT protein results suggest that liver function deteriorates with increasing abdominal obesity with VAT area positively associated with ALT, ALK and GGT. By contrast, the inverse relationship between visceral fat and bilirubin is consistent

with the inverse association between bilirubin, insulin resistance (Lin *et al.*, 2009) and GGT activity (Giral *et al.*, 2010). In particular, of the four tests, the linear association between ALT and adiposity is specifically mediated via visceral fat, while for ALK, BIL and GGT there is only evidence to suggest this is the case for abdominal fat *per se* (i.e. residuals analysis for ALK, BIL, and GGT demonstrates that the association between LFT and visceral fat and DXA total abdominal fat remains, conditional upon BMI, but not for association between LFT and visceral fat conditional upon BMI and DXA total abdominal fat). The visceral adipose depot is thought to be the major source of elevated fatty acid delivered to the liver via the portal vein and it is possible that visceral fat acts as a marker of hepatic fat content (Fabbrini *et al.*, 2009). Hepatic fat accumulation is known to impair insulin signalling in hepatocytes (Samuel *et al.*, 2010, Samuel *et al.*, 2004) and "pathway selective insulin resistance"(Brown and Goldstein, 2008) maybe pivotal in the transition from normal to impaired fasting glucose states. It has been hypothesised that NALFD may play a mediatory role in the pathology of CVD (Targher *et al.*, 2010).

The analysis of incident subclinical atherosclerosis is consistent with cross-sectional studies that show association between atherosclerosis, total abdominal fat (Kramer *et al.*, 2009) and visceral fat (Lear *et al.*, 2007, Kim *et al.*, 2009). Here we provide additional evidence that total abdominal and visceral fat accumulation precedes atherosclerosis.

The limitations of this study include that adiposity inferences are restricted to middle-aged Caucasian women. Not having sagittal depth for the study sample limited the ability to estimate visceral fat more accurately. This is indicated by the best anthropometric validation sample estimate (Model 2) predicting VAT area as well or better than a combination of DXA total abdominal fat, waist circumference and age (Model 1). As shown in the results, sagittal depth is an extremely useful yet simple anthropometric measurement to estimate the body cavity area, when used in combination with the subcutaneous fat width, gives a very good approximation of visceral fat ($R^2 \approx 90\%$). Where DXA is not available, sagittal depth, waist circumference and subcutaneous fat width could be measured to quickly and inexpensively estimate a person's visceral fat content with a tape measure and skin callipers.

Although the prediction of visceral fat was valid (O'Brien, 2007), co-linearity between model explanatory variables – in particular VAT area and DXA total abdominal fat – meant that adjusted odds ratios for risk factors were not interpretable. Instead, conditional independence structures were assessed using residuals analysis. This allowed us to eliminate the possibility of causal relationships between adiposity variables and morbidity, except those mediated by visceral fat.

The twin-based heritability estimate presented is likely to be an under-estimate, since random measurement error is included in the denominator of the heritability ratio estimate and the study sample estimate of visceral fat is known to account for approximately only 83% of CT-measured VAT area.

The study design for incident subclinical atherosclerosis was not strictly prospective, since no baseline measure of cIMT was taken to facilitate the exclusion of potential baseline cases. Where undetected baseline cases exist, the study design is cross sectional.

Although not measured the same day, time between DXA and CT scans visits (2 year maximum) was included in analyses as a confounding variable and Bland-Altman plots revealed no bias in the relationship between predicted and CT-measured visceral fat values.

Where a direct measure of visceral fat is unavailable, using an indirect estimate of visceral fat (even without a measure of subcutaneous fat) is more predictive of morbidity than total abdominal fat. This study suggests that both DXA body fat composition and anthropometric measures of body cavity volume alone can provide reliable estimates of abdominal visceral fat.

In conclusion, this study in middle-aged women demonstrates that a DXA-based measure of visceral fat is associated with T2D, hypertension, subclinical atherosclerosis and liver function. It is the predominant measure of adiposity risk for T2D, hypertension and alanine aminotransaminase serum level. This is consistent with hypotheses that suggest excess visceral fat accumulation is causally related to cardiovascular and metabolic morbidities.

Chapter 4: A Candidate Gene Study for the *PARL/ABCC5* Gene Region as a Novel Type 2 Diabetes Susceptibility Locus

Abstract

A candidate polymorphism and candidate gene study of a linkage disequilibria genomic region including *PARL* and *ABCC5* was conducted for European and African American samples, in relation to the phenotypes of fasting plasma insulin and glucose levels, visceral fat accumulation and T2D. Previous functional studies, animal models of T2D and human observational data have implicated *PARL*, with the latter including a report that a non-synonymous SNP in exon 7 of *PARL* (Val262Leu, rs3732581) may be a causal functional variant that confers disease risk.

No evidence of phenotypic association with *PARL* was found using commercial and fine map genomic data (European and African American samples). By contrast, strong evidence of phenotypic association within intron 26 of the *ABCC5* gene was observed for European and African American samples, which is located adjacent (35Kb) and in LD with *PARL*. The genomic location estimate for the *ABCC5* disease susceptibility variant associated with T2D-related phenotypes and expression data was the same for all samples (185,136Kb), with the location restricted to *ABCC5* (95% CI 185,108-185,235Kb). The expression data strongly suggests the identified genomic variant regulates *ABCC5* mRNA expression levels with *ABCC5* expression also associated with T2D-related phenotypes, including T2D.

Previous observations of human genomic association with *PARL* may be due to LD confounding with neighbouring *ABCC5*. The *ABCC5* variant is associated with *ABCC5* expression, fasting insulin, glucose and T2D, while *ABCC5* expression is itself an intermediate phenotype for T2D. Given the *ABCC5* risk variant is cosmopolitan, common and observed in populations of disparate ancestry, this indicates the polymorphism is old and that the transporter gene is likely to play an aetiological role in the onset of T2D

Introduction

The long arm of chromosome 3, specifically 3q26-29, has been gained interest in T2D research due to repeated evidence of linkage to this locus for glycaemic traits (Francke *et al.*, 2001) and T2D (Guan *et al.*, 2008, Mori *et al.*, 2002, Vionnet *et al.*, 2000, Hegele *et al.*, 1999, Walder *et al.*, 2005, Civitarese *et al.*, 2010). GWA studies have identified two genes (*ADIPOQ* and *IGF2BP2*) in this ~37Mb region that are associated with T2D, but given the replicated linkage for T2D and T2D-related traits, there are likely to be many additional disease susceptibility genes to be found at this locus with a more common allelic spectra.

As part of a gene expression study designed to identify NEM genes associated with T2D, Walder *et al.* (2005) screened differentially expressed genes across the genome using skeletal muscle from hyper- and normo-glycaemic Israeli sand rats (*Psammomys obesus*). The list of differentially expressed genes were then filtered by retaining only genes that were thought to have mitochondrial roles and located in regions thought to harbour obesity and T2D susceptibility loci, only one gene met this criteria, *PARL*, located in 3q27. The authors go on to demonstrate that the expression of this NEM gene is reduced in insulin resistant and T2D-induced sand rats, but can be restored with an exercise regime of the diabetic rats.

In humans, a positive linear correlation was detected between *PARL* expression skeletal muscle and insulin sensitivity (as assessed by glucose disposal during a euglycaemic clamp). *PARL* expression and mtDNA copy number are reduced in T2D individuals compared to controls and both measures decline with age (Civitarese *et al.*, 2010). Similarly, insulin plasma levels are raised in rats fed a high fat diet compared to those fed a standard diet leading to a reduction in *PARL* expression and mitochondrial function (Tang *et al.*, 2009). Genetic polymorphisms in *PARL* are associated with mitochondrial content (Curran *et al*., 2010) and it is plausible that variation in mitochondrial function may predispose individuals to develop overt T2D (Liesa *et al.*, 2009, Newsholme *et al.*, 2012, Patti and Corvera, 2010, Szendroedi *et al.*, 2012).

Based upon a dominant model, Walder *et al.* (2005) also observed a non-synonymous SNP (Val262Leu/rs3732581) in the seventh exon of *PARL* to be associated with

fasting insulin in 1,031 human subjects with the variant interacting strongly with age, in which fasting insulin levels for GG homozygotes were greater than GC/CC carriers with increasing age. While there is suggestive evidence of earlier age of T2D onset for GG carriers (Hatunic *et al.*, 2009), the SNP-insulin association has not subsequently been replicated (Hatunic *et al.*, 2009, Fawcett *et al.*, 2006, Powell *et al.*, 2008). Proteins that regulate the dynamic process of mitochondrial turnover (mitochondrial fusion and fission) are believed to play an essential role in quality control, disposing of dysfunctional mitochondria and maintaining collective functioning of mitochondria as a healthy "super-organelle" (Poole *et al.*, 2010, Suen *et al.*, 2008, Chan, 2006, Scorrano, 2007).

It is therefore of great interest that PARL, one of the mitochondrial proteases that is involved in the degradation of the key mitochondrial fusion protein OPA1 (Pellegrini and Scorrano, 2007), may also be a genetic risk factor for insulin resistance and T2D. Although the relationship is not understood, there is evidence to suggest that mitochondrial fusion activity is a defective process under insulin-resistant states (Liesa *et al.*, 2009).

The aims of this study are: 1) to confirm if the *PARL* missense Val262Leu polymorphism is associated with fasting plasma insulin levels and 2) to conduct a comprehensive candidate gene study of the *PARL*/*ABCC5* gene region within an extended LD block covering the approximate physical distance of 185,000-185,250Kb (build 36), to test for evidence of genetic of association with disease and glycaemic traits.

Given the strong functional evidence implicating *PARL* and the fact that the Val262Leu polymorphism is predicted to be benign (Adzhubei *et al.*, 2010), this study examines the hypothesis that the contradictory human genetic association results observed for Val262Leu may be due to confounding LD between this polymorphism and the true functional variant(s) located elsewhere within the gene or gene region.

To address these aims, two T2D case control samples with available genomic data (WTCCC1 European and NIDDK African American samples, see Chapter 2) and a population-based sample of healthy European twins (TwinsUK) were used. The latter

also measured for fasting insulin and glucose plasma levels, commercial array genomic data, additional fine mapped genotypes (including Val262Leu) and mRNA expression data for the *PARL* and *ABCC5* genes.

Materials and Methods

Glycaemic Traits

Insulin and glucose values were available for 5,605 TwinsUK subjects. Residuals for fasting glucose were taken by regression upon year of visit (categorical variable), age and sex (R^2 = 0.18); residuals for fasting insulin were carefully taken separately for each the three assays, by regressing each assay upon year of visit (categorical variable), age and sex (R^2 = 0.07, 0.07 and 0.06, respectively) and then residuals combined. Outliers for glucose residuals regressed upon insulin residuals were identified and removed (n = 103) using leverage and residual diagnostics. Visual inspection and statistical assessment of insulin residuals by year of collection showed no evidence of residual mean differences or heteroscedascity. The fasting plasma IGR was calculated for the remaining residuals and then quantile normalised. IGR was used for analysis to be consistent with Kissebah *et al.* (2000) and to have one summary glucose homeostasis variable for analysis. T2D status was ascertained via an online questionnaire asking if a physician had ever diagnosed the individual with the condition. These individuals (n = 338) were removed for analyses of healthy variation, leaving 5,164 individuals with valid fasting plasma insulin and glucose measures. HOMA2 values were calculated (Wallace *et al.*, 2004) for beta cell function (HOMA2-%B) and insulin sensitivity (HOMA2-%S).

Fine-Mapping

The Val262Leu (rs3732581) SNP was not present on either the HumanHap300 or Human610-Quad commercial arrays and was therefore genotyped for a total of 3,087 individuals by KBioscience (http://www.kbioscience.co.uk). For TwinsUK sample with HumanHap300 data, the coverage within the *PARL*/*ABCC5* gene region was poor so an additional 23 SNPs were genotyped to provide extra coverage and improved resolution of the LDU steps in this region. For TwinsUK sample with Human610-Quad data, an additional 26 fine-map SNPs were genotyped across the analytical window (184,743-185,548Kb, build 36) with a wide range of minor allele frequencies (MAF > 0.05) to further improve coverage for the *PARL* and *ABCC5* genes, as assessed using the CEU LDU map.

Statistical Genetic Methods

Biometric Model

For the candidate marker Val262Leu test of association, the quantitative traits of fasting insulin, glucose and IGR were regressed upon the SNP using a biometric additive model (Chapter 2: Biometric model). For the biometric additive model, the three genotypes for the SNP rs3732581 were coded G/G (0), G/C (1) and C/C (2) and utilised as an ordinal variable with the regression coefficient interpreted in a minor allele, dose-dependent manner. For example, an additive SNP model for fasting insulin plasma levels (if significant) would be interpreted as pmol/l decrease per copy of minor allele. Similarly, to construct the intermediate file of summary results to implement the Malécot test of association, the phenotype was regressed upon each SNP using an additive biometric model.

An interaction model testing for association between fasting insulin levels and age stratified by Val262Leu genotypes was implemented using a nested interaction design. A mixed linear regression model was implemented to regress fasting plasma insulin upon two age terms, one for GG individuals (age_{GG}) and another age variable for GC/CC individuals ($age_{GC/CC}$). Fixed effects were included in the regression for a dummy intercept term to identify the two genotype strata, year of visit (categorical variable) and a random effects intercept term using family identifier. Familial relatedness, i.e. non-independence between individuals, was incorporated into the analysis by clustering individual observations by family identifier.

eQTL Regression Analyses

Gene expression data in three tissues (subcutaneous adipose tissue, skin, and LCL) for *PARL* and *ABCC5* were available for a subset of the TwinsUK sample (n = 821); the generation of this data is described in Chapter 2. There were three probes per gene, each probe tagging a mixture of mRNA transcripts; the location and the associated transcripts for each probe are provided in Supplementary Table 4.1

To control for observed systematic experimental protocol effects (different means and variances for the three batch groups), mixed linear regression methods were used to model normalised expression probe values (18 in total, 3 probes per gene and from 3

tissues). Each $\log_2$ normalised expression probe was regressed upon SNP plus fixed effect confounder variables of age at sample collection and sample batch (a 3-level factor). While the regression slope terms were fixed, the intercept included an additional random effects term for the twin family identifier to estimate variance between families in order to be able to make more reliable population inferences.

Drs Winston Lau and Nikolas Maniatis generated summary statistics from the SNP regression models for the entire genomic region (184,743-185,548Kb, build 36) and fitted a Malécot model to test for evidence of regulatory eQTLs in the region for each gene probe.

Multiple Testing and Significance Threshold

For the Malécot test of association, which is a single test, the significance threshold is an alpha of 5%. For the eQTL analyses, in which a total of 18 tests were performed (two genes, three probes per gene across three tissues), a Bonferroni corrected alpha threshold of $3x10^{-3}$ (0.05/18) was used.

Linkage Disequilibrium Maps and Genetic Association

There are three major strengths to a genetic map-based test of association. The first is that this elegant model is highly interpretable in terms of the ancestry and age of the susceptibility variant. The second is that the model does not assume that the disease susceptibility variant has been genotyped, nor that it is in high or complete LD with one of the genotyped markers, which is a major assumption in standard GWA scans. By utilising genetic maps, the model is sensitive to detecting disease susceptibility variants that are potentially associated with a constellation of neighbouring genotyped markers, but not necessarily in high LD with any one individual genotyped marker. For the commercial SNP arrays that are currently available and where the assumption of high allelic identity (i.e. a single disease-susceptibility allele at a locus) begins to breakdown (Pritchard *et al.*, 2000), this is a much more realistic scenario than the current practice of assuming a variant is in high LD with one of the genotyped markers. Direct estimation of the disease susceptibility variant location also provides a degree of commensurability between different commercial genotyping platforms that other methods do not provide. Thirdly, gene mapping is far more efficient using markers located upon a genetic rather than a physical map (Maniatis *et al.*, 2004).

Drs Winston Lau and Nikolas Maniatis generated all LDU maps and applied the muli-marker Malécot test of association. They provided me with their summary results that I used for analysis in Chapters 4 and 5 of this thesis.

Results

A description for each of the three samples (TwinsUK, WTCCC1, and African American) is provided in Chapter 2. Summary statistics of the three study samples are provided in Table 4.1. European and African American data sets each contained over 1,000 cases of T2D with approximately equal (African American) or greater numbers of controls (European) and equal numbers of men and women in both groups. While the European and African American case-control samples for T2D provided a test of genetic association with overt disease, the TwinsUK sample, for which T2D cases were excluded from the analysis of glycaemic traits, provided a test of association with healthy variation in fasting insulin and glucose plasma levels. All samples were middle-aged ranging from approximately 40-70 (WTCCC1 and NIDDK) and 20-80 (TwinsUK) years old.

	Europeans Affymetrix 500K		**African Americans** Affymetrix 6.0		**TwinsUK**	
	Cases	*Controls*	*Cases*	*Controls*	Human Hap300	Human 610-Quad
Mean IGR	-	-	-	-	11.2	11.1
Mean age	58.3	45.4	61.5	49.0	50.7	50.6
Female %	42	51	57.3	61.2	99.9	88.0
n subjects (array)	1,924	2,938	1,033	971	1,750	2,300
n subjects (fine-map)	-	-	-	-	801	2,300

Table 4.1. Summary statistics for the three study samples used for the *PARL*/*ABCC5* candidate gene study.

Array refers to the number of individual with genome-wide SNP data generated on commercial arrays. Where there was poor coverage for the TwinsUK sample relative to European and African American, additional SNPs were genotyped in TwinsUK sample with genome-wide SNP data (fine map). Abbreviations: IGR - insulin:glucose ratio (pmol/L per mmol/l).

Val262Leu (rs3732581, *PARL*) as a Candidate Marker for Insulin Resistance and T2D

The rs3732581 non-synonymous polymorphism in *PARL* was not included in the genotyping platforms (Table 4.1). Therefore this marker was genotyped for a total of 3,087 individuals from the TwinsUK sample, providing complete genotype and phenotype fasting insulin data for 2,358 healthy non-T2D individuals and 2,895 T2D case-control data (n_{cases} = 132). The observed minor allele frequency (MAF) was 0.489 (allele C), which is similar to in documented European, African and Asian populations (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=3732581).

A previously suggested dominant mode of inheritance for Val262Leu implies that GC and CC carriers are likely to witness smaller increases in fasting insulin with age compared to individuals with the GG genotype at this locus (Walder *et al.*, 2005). Attempting to replicate this age by genotype interaction model for Val262Leu, both for fasting insulin and T2D, phenotypes were regressed upon two age terms in the same model, in order to test whether the strength of association with age for GG individuals (age_{GG}) was stronger than those carrying GC/CC Val262Leu genotypes ($age_{GC/CC}$, Chapter 4: Materials and Methods).

Mixed linear regression models showed no evidence of association (unadjusted or adjusted for age) for fasting plasma insulin levels (p-value = 0.95) or self-reported T2D (p-value = 0.10) with the Val262Leu polymorphism parameterised using a simple additive biometric model (Chapter 4: Materials and Methods). By contrast, the same regression models showed strong statistical evidence that fasting insulin levels (0.12pmol/L per year) and risk of T2D (OR = 1.04 per year) increased with age (Table 4.2A).

Although both phenotypes were positively associated with age, GG individuals showed no evidence of stronger association between phenotype and age compared to CC/GC individuals (Table 4.2B), with age regression coefficients $b_{GG} = b_{GC/CC}$ for insulin (Wald $\chi^2_1 = 0.14$, p-value = 0.71) and $b_{GG} = b_{GC/CC}$ for T2D ($\chi^2_1 = 0.74$, p-value = 0.39).

Model		Fasting Insulin			Type 2 Diabetes		
		β	SE	p-value	OR	SE	p-value
(A)	rs3732581	-0.17	0.56	0.95	0.82	0.10	0.10
	Age	0.12	$3x10^{-2}$	$1x10^{-4}$	1.04	0.01	$3x10^{-8}$
(B)	$Age_{G/G}$	0.13	0.06	0.03	1.03	0.01	0.01
	$Age_{G/C\&C/C}$	0.12	0.04	$1x10^{-3}$	1.05	0.01	$3x10^{-9}$

Table 4.2. Non-synonymous SNP rs3732581 (Val262Leu) association with fasting plasma insulin levels and type 2 diabetes for TwinsUK sample.

(A) Main effect model: association between phenotype and SNP adjusted for age, with rs3732581 coded 0, 1, 2 copies of the C allele for an additive biometric model. **(B)** Interaction model association between phenotype and age stratified by SNP using a nested interaction model adapted from Walder *et al.* (2005), in which age is stratified into two genotype groups with the C allele completely dominant over G. Fasting plasma insulin levels measured in pmol/L. Year of blood sample collection was included as a categorical confounding variable for both fasting insulin analyses. Sample sizes were 2,358 and 2,895 for insulin and T2D analyses, respectively.

LD in the *PARL/ABCC5* Gene Region

Figures 4.1 and 4.2 illustrate the high degree of LD as measured by D' (see Chapter 2 for details on LD measures) within and moderate LD between the *PARL* and *ABCC5* genes, with Figure 4.1 plotting pair wise LD and Figure 4.2 plotting the informative fine scale genetic maps for HapMap European (CEU) and African (ASW) populations in the form of cumulative LDU. Note that while there is breakdown in LD between the two genes, there still remains some extended LD between the two. The mapping approach used in this study uses genetic locations derived from LD maps with distances in LDU inferred using population-specific HapMap genomic data (Maniatis *et al.*, 2002). The maps can be visualised by plotting marker location in LDU against the marker physical distances in kilobases (Kb). By plotting these LDU maps against physical distance, the non-linear relationship is revealed as a "Block-Step" structure (Figure 4.2). "Blocks" of LD represent areas of low haplotype diversity, while "steps" define LD breakdown, mainly caused by recombination. The high-resolution genetic maps (based upon HapMap markers) provide detailed local LD structure and can be used for mapping potential functional variants associated with fasting insulin and glucose plasma levels in healthy Europeans (TwinsUK) and T2D for European and African American samples.

Given the lack of association between Val262Leu with fasting insulin and T2D, the investigation was extended to a more comprehensive candidate gene study of the region using genomic data for three samples (Table 4.1). Additional SNPs were genotyped for the TwinsUK HumanHap300 samples to make the markers for these samples more compatible with the Human610-Quad coverage in the *PARL/ABCC5* region (Chapter 4: Materials and Methods). For the candidate gene association study the phenotypes used were fasting IGR (TwinsUK) and T2D status (WTCCC1, NIDKK and TwinsUK).

Figure 4.1. Haploview LD plot illustrating extended linkage disequilibrium in the *PARL/ABCC5* region.

Pair-wise LD (r^2) from Haploview (Barret *et al.*) using HapMap3 (release 2) for the CEU population.

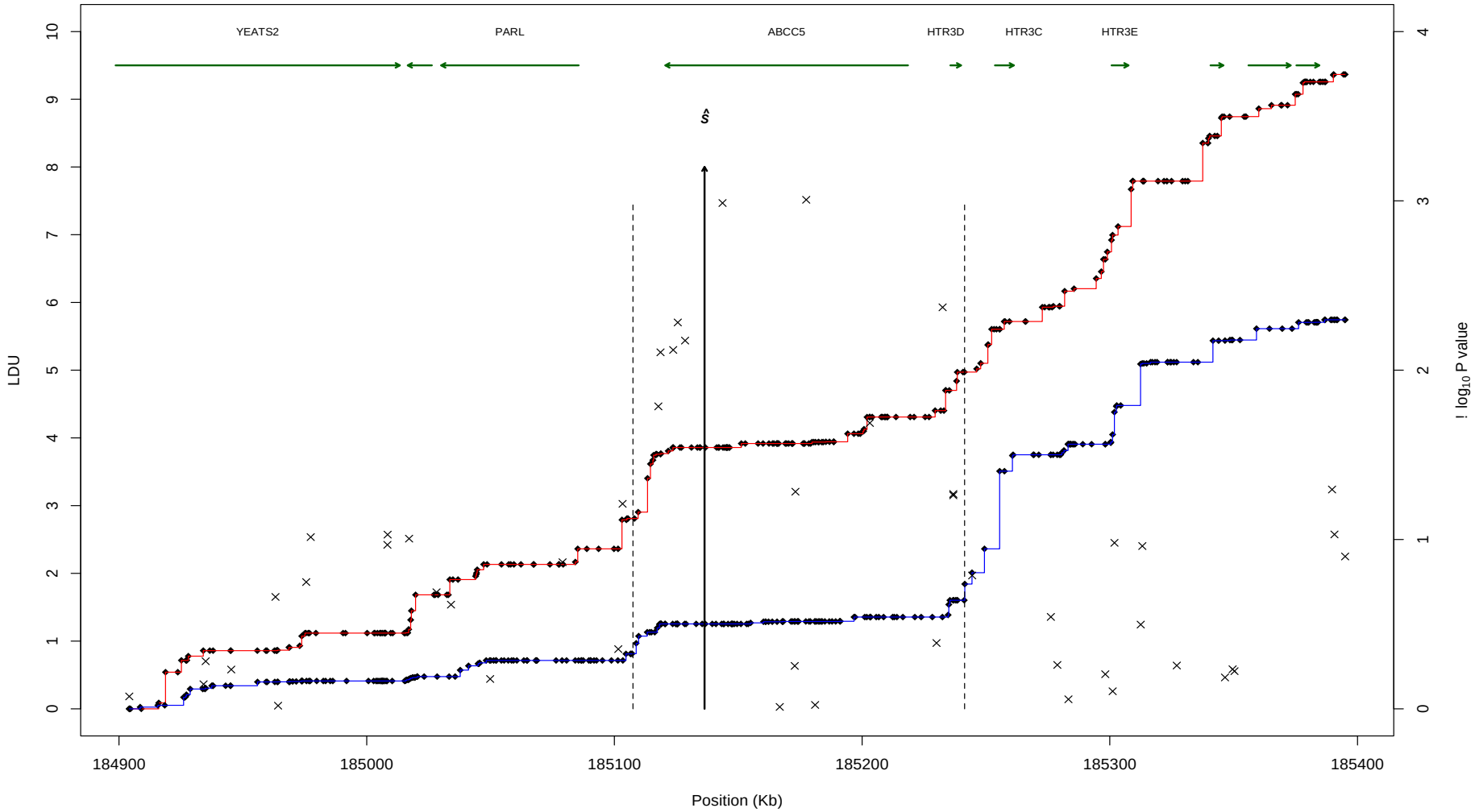

YEATS2
PARL
ABCC5
HTR3D
HTR3C
HTR3E
LDU
! log10 P value
Position (Kb)
184900
185000
185100
185200
185300
185400
0
1
2
3
4
5
6
7
8
9
10

Figure 4.2. African American and European genetic maps and scatter plot for WTCCC1 association between type 2 diabetes and single nucleotide polymorphisms.

The graph presents a line plot for African American (red) and European (blue) cumulative LDU genetic maps (y1-axis) and scatter plot for WTCCC1 T2D association with SNPs in the *PARL/ABCC5* gene region (y2-axis, $-\log_{10}$ p-value). The genetic maps for HapMap Africans (ASW, Phase III) and Europeans (CEU, Phase II) are plotted, with horizontal regions reflecting "blocks" or extended local LD, while vertical steps, reflect breakdown in LD. Each dot on these lines represents a SNP location for the high-resolution HapMap samples from which these population-specific genetic maps are inferred. The arrowed vertical line labelled $\hat{S}$ represents the functional variant location estimate and the dotted lines the 95% CI for the variant location. The total genetic distance (y1-axis) for the same physical genomic region (x-axis) is greater for the African population compared to the European, reflecting older ancestry and larger number of historical recombination events for this population. Note that while none of the individual SNP p-values for WTCCC1 are smaller than 0.001 ($-\log_{10}[0.001] = 3$), collectively they provide strong evidence of association with T2D (Malécot model p-value = 3×10^{-6}). The nearest SNP to the $\hat{S}$ location (arrow) is rs4912515.

The *PARL*/*ABCC5* Gene Region as a Candidate for Insulin Resistance and T2D

Under the supposition that there is at least one detectable common functional polymorphism (the "common disease, common variant" or CDCV hypothesis (Pritchard, 2001, Reich and Lander, 2001) associated with both insulin sensitivity and T2D and that there is minimal allelic heterogeneity in the gene region (Bodmer and Bonilla, 2008, Pritchard and Cox, 2002), we examined the *PARL*/*ABCC5* region in depth by employing a powerful multi-marker test of association (see Chapter 4 Materials and Methods). This approach uses information from all SNPs within a genomic region (analytical window) and utilises the genetic distances between markers from fine scale, population-specific LDU maps. The method provides a location estimate, $\hat{S}$, (and 95% CI or location support interval) for the putative functional common variant, if there is significant evidence of phenotypic association in a genomic region of interest.

Data	Phenotype	Cohort	n	SNPs	p-value	$\hat{S}$ (Kb)	Lower 95% CI	Upper 95% CI	Nearest gene (Transcript strand)
Genomic	T2D	WTCCC1 T2D	4,864	83	$3x10^{-6}$	**185,136**	185,107	185,241	Intron 26 *ABCC5* (-)
Genomic	T2D	NIDDK T2D	1,937	104	0.01	**185,137**	185,114	185,218	Intron 26 *ABCC5* (-)
Genomic	IGR	TwinsUK Human610-Quad	1,700	176	0.04	**185,136**	185,110	185,235	Intron 26 *ABCC5* (-)
Genomic	IGR	TwinsUK HumanHap300	801	72	0.50	-	-	-	-
Genomic meta-analysis					$1x10^{-6}$	**185,136**	185,108	185,235	Intron 26 *ABCC5* (-)
Subcutaneous adipose expression	*ABCC5* mRNA	TwinsUK	276-700*	202	$1x10^{-11}$	**185,136**	185,109	185,235	Intron 26 *ABCC5* (-)

Table 4. 3. Phenotype-genotype association for *PARL*/*ABCC5* gene region in Europeans and African Americans.

WTCCC1 showed significant evidence of association between T2D and *ABCC5* with a causal variant ($\hat{S}$) estimated to be located in intron 26. This is replicated for healthy Europeans measured for fasting insulin and glucose (TwinsUK Human610-Quad, p-value = 0.04) and African American samples (p-value = 0.01) with three data sets providing the same variant location at or close to 185,136Kb (95% CI: 185,108-185,235Kb). Meta-analysis of the four genomic sample data sets including both T2D and IGR showed evidence for an intragenic functional variant in *ABCC5* with a p-value of $1x10^{-6}$. Subcutaneous adipose expression of transcripts tagged by the Ilmn_1706531 probe showed strong evidence (p-value = $1x10^{-11}$) that the intron 26 variant is an eQTL that regulates *ABCC5* expression levels. Analytical window co-ordinates used for phenotype and transcript analyses were: 184,743-185,548Kb (build 36). *TwinsUK sample size for individuals with both expression and SNP data ranged from 276 to 700 for a total of 202 SNPs in the *PARL*/*ABCC5* gene region, with samples genotyped for SNPs on both the Human610-Quad and HumanHap300 arrays (72 SNPs, n = 700), the Human610-Quad only (104 SNPs, n = 276) and 26 additional fine map SNPs for the *PARL*/*ABCC5* genes (n = 276). Sample size for each SNP was taken into account for all analyses. Abbreviations: IGR - fasting insulin: glucose serum ratio levels, NIDDK – National Institute for Diabetes, Digestive and Kidney disease, T2D - type 2 diabetes, WTCCC1 - Wellcome Trust Case Control Consortium.

Table 4.3 presents evidence for phenotypic association with genomic polymorphism data centred on the *PARL*/*ABCC5* gene region (analytical window: 184,743-185,548Kb, build 36) for European and African American populations. The analytical window size is approximately 10 LDU in genetic distance (for European populations), which is an arbitrarily chosen size, but empirically large enough to be informative (Chapter 2: Linkage Disequilibrium). The WTCCC1 T2D sample had a total of 83 SNPs genotyped in this region (Affymetrix 500K) for a total of 4,864 individuals, with significant evidence of association (p-value = $3x10^{-6}$) observed for a location estimate at 185,136Kb (95% CI 185,108-185,215Kb). Note that while the best evidence for the location of the common functional variant is intron 26 of *ABCC5*, the confidence interval for the location estimate includes the entire *ABCC5* gene.

Interestingly, an essentially identical location estimate, 185,137Kb (95% CI: 185,114-185,218Kb), was obtained for African American T2D case-control sample with nominal significance (p-value = 0.01). The result is also replicated for healthy Europeans (TwinsUK Human610-Quad) measured for IGR, the analysis yielded a p-value of 0.04 for the same location estimate as T2D and within *ABCC5* intron 26 (185,136Kb; 95% CI 185,110-185,235Kb). The lack of association for the TwinsUK HumanHap300 sample maybe due to poorer marker coverage in the gene region (a total of 72 SNPs) compared to the other commercial SNP arrays. Meta-analysis of the p-values from the four genomic results (Table 4.3; WTCCC1 p-value = $3x10^{-4}$, NIDDK p-value = 0.01, TwinsUK Human610-Quad p-value = 0.04 and TwinsUK HumanHap300 p-value = 0.50) using Fisher's method (Fisher, 1925), nevertheless provided overall significant evidence of association (p-value = $1x10^{-6}$) for a common functional variant located at 185,136Kb. The 95% confidence intervals for this location range from 185,108-185,235Kb, which includes the *ABCC5* gene, promoter and 3' region (*ABCC5* gene co-ordinates: 185,120-185,215Kb).

Using expression data for *PARL* and *ABCC5*, an expression quantitative trait locus (eQTL) analysis was performed and the summary p-values were fitted to the Malécot model in order to obtain location estimates across multiple tissues. Table 4.3 shows the association results for one of the *ABCC5* subcutaneous adipose expression probes (Ilmn_1706531) measured for the TwinsUK sample, the expression of its transcripts is associated (p-value = $1x10^{-11}$) with T2D and provides the same location estimate as

the genomic tests (185,136Kb; 95% CI: 185,109-185,235Kb). This result strongly suggests that the associated functional variant in *ABCC5* is an eQTL that regulates *ABCC5* transcript expression levels.

Tissue	Gene	Probe	p-value	$\hat{S}$ Kb (B36)	Lower 95% CI (Kb)	Upper 95% CI (Kb)	eQTL location
Subcutaneous adipose tissue	*PARL*	Ilmn_2341467	$4x10^{-1}$	-	-	-	-
	PARL	Ilmn_1731354	$2x10^{-1}$	-	-	-	-
	PARL	Ilmn_2257665	$2x10^{-1}$	-	-	-	-
	ABCC5	Ilmn_1706531	$1x10^{-11}$	**185,136**	**185,109**	**185,235**	*ABCC5 intragenic*
	ABCC5	Ilmn_1651964	$9x10^{-2}$	-	-	-	-
	ABCC5	Ilmn_2302358	$2x10^{-2}$	-	-	-	-
LCL	*PARL*	Ilmn_2341467	$4x10^{-13}$	185,111	185,108	185,192	-
	PARL	Ilmn_1731354	$2x10^{-8}$	185,109	185,104	185,118	3' *ABCC5 intra/intergenic*
	PARL	Ilmn_2257665	$4x10^{-5}$	185,302	185,301	185,307	3' *ABCC5 intergenic*
	ABCC5	Ilmn_1706531	$9x10^{-22}$	185,179	185,117	185,235	*HTR3E intragenic*
	ABCC5	Ilmn_1651964	$2x10^{-54}$	185,193	185,118	185,235	*ABCC5 intragenic*
	ABCC5	Ilmn_2302358	$8x10^{-4}$	185,421	184,743	185,421	*ABCC5 intragenic*
Skin	*PARL*	Ilmn_2341467	$4x10^{-5}$	185,076	184,928	185,117	-
	PARL	Ilmn_1731354	$1x10^{-12}$	185,109	185,104	185,117	-
	PARL	Ilmn_2257665	$3x10^{-2}$	-	-	-	*PARL intragenic*
	ABCC5	Ilmn_1706531	$9x10^{-22}$	185,170	185,116	185,235	5' *PARL intergenic*
	ABCC5	Ilmn_1651964	$1x10^{-6}$	185,179	185,110	185,235	-
	ABCC5	Ilmn_2302358	$1x10^{-12}$	185,104	185,040	185,108	*ABCC5 intragenic*

Table 4.4. Expression quantitative trait locus (eQTL) association analysis for TwinsUK data across three tissues.

The eQTL location estimate ($\hat{S}$) for *PARL* and *ABCC5* expression values in subcutaneous adipose, LCL and skin tissues are presented (n = 821). Each analysis is based upon the same analytical window co-ordinates used for the genomic analysis (184,743-185,548Kb, see Table 4.3). The p-value from the Malécot association model represents the statistical evidence for the observed probe-SNP data being consistent with the null hypothesis of no association for the entire gene region. Only one of the six probes for subcutaneous adipose tissue (Ilmn_1706531) showed evidence for an eQTL - located in intron 26 of *ABCC5* (185,136Kb, 95% CI 185,109-185,235Kb) - that regulates *ABCC5* expression levels. By contrast, most of the expression probes for LCL and skin tissues showed evidence of eQTL regulation in the same or neighbouring gene to which the probe was located. In LCL tissue, distal cis-acting eQTLs located in *ABCC5* and *HTR3E* appears to regulate *PARL* expression and in skin, an eQTL located in *PARL* regulates *ABCC5* expression levels. A conservative Bonferroni correction threshold of $\chi^2_1 = 10$ (p-value = 0.002) was used to control the Type I error rate for the 18 tests. All genomic locations are build 36 (B36).

Gene Expression Quantitative Trait Locus (eQTL) Analysis

The results from the subcutaneous adipose expression data in Table 4.3 were based on the *ABCC5* subcutaneous adipose expression probe (Ilmn_1706531), since this was the only adipose probe that showed evidence of genetic association. Table 4.4 lists the results for the eQTL analysis of all *PARL* and *ABCC5* mRNA probes, measured in subcutaneous adipose, LCL and skin tissues for a subset of the TwinsUK sample (Chapter 2: TwinsUK Sample). Eighteen expression probes were tested using mixed linear regression methods (Chapter 4: Materials and Methods) – three probes in each gene measured for three tissues. Of the three tissues most relevant to T2D, subcutaneous adipose tissue showed evidence (p-value = $1x10^{-11}$) for genetic association with only one probe (Ilmn_1706531, *ABCC5*), with the eQTL having exactly the same *ABCC5* (intron 26) location estimate as the functional variant location estimates for the two associated phenotypes, T2D and IGR. None of the other five subcutaneous adipose tissue transcript probes showed evidence of *cis*- genetic association (Table 4.4).

As noted, the location estimates for genomic and subcutaneous adipose expression data allude to the same functional variant in *ABCC5*. Although the *ABCC5* variant has not been genotyped, by way of illustration, the nearest genotyped SNP to the functional variant location estimate is rs4912515 (Figures 4.1 and 4.2). The regression coefficient for subcutaneous adipose probe Ilmn_1706531 regressed upon rs4912515 using an additive model is -0.07 (p-value = 0.003), which is equivalent to approximately 0.1 of a standard deviation decrease in *ABCC5* expression levels for each G allele inherited at this locus (MAF = 0.40). Except for adipose probe Ilmn_1706531, SNP rs4912515 is not associated with any of the other transcript probes (for all three tissues) or phenotypes (data not shown). SNPs in immediate proximity to the variant location estimate in *ABCC5* intron 26 showed evidence of long range LD. For example, pair wise LD between Val262Leu and rs4912515 (185,136.7Kb) is D' = 0.7 and r^2 = 0.31 (Figure 4.1). The MAF for Val262Leu (rs3732581) and rs4912515 SNPs are similar and common in Europeans (0.49 and 0.40, respectively), Asians (CHB = 0.47, 0.40) and Africans (YRI = 0.48, 0.40).

In addition to the Val262Leu SNP in *PARL*, another SNP in *PARL* (rs3792588) has been reported to be associated with mitochondrial DNA content (Curran *et al.*, 2010); this SNP (TwinsUK, MAF = 0.14) is in LD with Val262Leu (D' = 1 and r^2 = 0.13, Figure 4.1). This raised the possibility of a link between the Val262Leu SNP-insulin association and mitochondrial abundance, however, rs3792588 showed no evidence of association with IGR or T2D (data not shown).

The expression results for LCL and skin tissues were more numerous than subcutaneous adipose and the pattern of association more complicated. *ABCC5* expression levels appeared to be mainly regulated by intragenic eQTLs. *PARL* expression levels were either regulated by eQTLs in neighbouring genes (*ABCC5* and *HTR3E* for LCL tissue) or by eQTLs upstream proximal to *PARL* (skin). In short, *ABCC5* transcripts showed evidence of "self" regulation for all three tissues, while for *PARL* there was (tissue specific) evidence of co-expression between genes. Table 4.5 presents a correlation matrix for all probes across the three tissues and Supplementary Table 4.1 lists the different splice variant transcripts for *PARL* and *ABCC5* that are targeted by the three probes in each gene.

The *ABCC5* Ilmn_1706531 probe is perhaps the most interesting because despite only tagging two transcripts, it is the only probe to be correlated with the all three *PARL* probes across the three tissues. Excluding correlations with the Ilmn_1706531 probe, the only *PARL* probe correlated with *ABCC5* is the Ilmn_2257665 probe, which is negatively correlated (r = -0.21) with Ilmn_1651964 in both subcutaneous adipose tissue and skin.

Tissue	Gene	Probe	**Ilmn_2341467**	**Ilmn_1731354**	**Ilmn_2257665**	**Ilmn_1706531**	**Ilmn_1651964**	**Ilmn_2302358**
Subcutaneous adipose tissue	*PARL*	Ilmn_2341467	1					
	PARL	Ilmn_1731354	0.81	1				
	PARL	Ilmn_2257665	-0.45	-0.36	1			
	ABCC5	Ilmn_1706531	0.28	0.23	-0.38	1		
	ABCC5	Ilmn_1651964	-	-	-0.21	0.54	1	
	ABCC5	Ilmn_2302358	-	-	-	-	-	1
LCL	*PARL*	Ilmn_2341467	1					
	PARL	Ilmn_1731354	0.88	1				
	PARL	Ilmn_2257665	-0.26	-0.23	1			
	ABCC5	Ilmn_1706531	0.11	0.08	-0.33	1		
	ABCC5	Ilmn_1651964	-	-	-	0.51	1	
	ABCC5	Ilmn_2302358	-	-	-	0.09	0.12	1
Skin	*PARL*	Ilmn_2341467	1					
	PARL	Ilmn_1731354	0.85	1				
	PARL	Ilmn_2257665	-0.42	-0.33	1			
	ABCC5	Ilmn_1706531	0.33	0.24	-0.49	1		
	ABCC5	Ilmn_1651964	-	-	-0.21	0.56	1	
	ABCC5	Ilmn_2302358	-	-	-	-	-	1

Table 4.5. TwinsUK transcript expression correlation structure for *PARL* and *ABCC5*.

The Pearson product-moment correlation is presented between six mRNA expression probes for *PARL* and *ABCC5* measured in adipose, LCL and skin tissue samples. The transcript correlations are for probes within the same gene, while highlighted areas refer to between-gene correlations (for the same tissue). The residuals for batch effects were first taken for each probe before the correlation coefficients were estimated to provide a partial correlation coefficient controlling for batch effect. Hyphen indicates a non-significant correlation (p-value > 0.05).

Subcutaneous Adipose *ABCC5* Expression as an Intermediate Phenotype for Fasting Insulin, Visceral Fat Accumulation and T2D

To assess if subcutaneous adipose *PARL* and *ABCC5* transcripts may be related to T2D and related phenotypes, a series of analyses were performed regressing each phenotype upon each of the six subcutaneous adipose probes in turn, using a mixed linear regression model to control for batch and age effects (Chapter 4: Materials and Methods). Visceral fat was included as a T2D-related phenotype as this phenotype has previously been shown to be strongly associated with T2D and arguably may play a causal role in T2D disease onset (Chapter 3). In particular, we were interested in adipose probe Ilmn_1706531 as this probe already showed evidence of eQTL genetic association with *ABCC5* (Table 4.3).

The univariate regression results presented in Table 4.6A indicates that only *ABCC5* subcutaneous adipose probe Ilmn_1706531 was significantly associated with fasting insulin, visceral fat accumulation and T2D. To control for correlation between expression probes, we also performed a multiple regression (Table 4.6B) in which phenotype was regressed upon Ilmn_1706531, age, batch effects and the two probes (Ilmn_2257665, Ilmn_1651964) that showed marginal evidence of phenotypic association in univariate analyses. This confirmed that only subcutaneous adipose probe Ilmn_1706531 was significantly (and strongly) associated with fasting insulin, visceral fat and T2D. The strength of association for this *ABCC5* probe (Table 4.6B) was estimated to have a correlation coefficient of 0.37 with IGR (95% CI 0.13 - 0.61), a regression coefficient of $30cm^2$ visceral fat per SD increase in expression (95% CI 13.2 - 47.4) and for T2D, an OR of 3.8 per SD increase in probe expression (95% CI: 1.25 - 11.6).

The association between IGR and subcutaneous adipose probe Ilmn_1706531 appears to be primarily driven by fasting insulin (p-value = $1x10^{-4}$) rather than glucose levels (p-value = 0.02) with raised subcutaneous adipose *ABCC5* transcript levels positively associated with beta cell function (homa2b, p-value = $1x10^{-3}$) and negatively with peripheral insulin sensitivity (homa2s, p-value = $5x10^{-5}$). Significantly, none of the expression probes for LCL and skin tissues – including *ABCC5* probe Ilmn_1706531 - were associated with IGR, visceral fat or T2D (data not shown).

	Gene	Probe	Fasting IGR			Visceral fat			T2D		
			β	SE	p-value	β	SE	p-value	β	SE	p-value
A	*PARL*	Ilmn_2341467	-0.01	0.18	0.96	-1.8	12.1	0.88	-0.71	1.02	0.48
	PARL	Ilmn_1731354	0.06	0.18	0.73	-13.9	12.8	0.28	-0.52	1.13	0.65
	PARL	Ilmn_2257665	-0.04	0.10	0.67	-15.1	7.5	0.04	-0.57	0.55	0.30
	ABCC5	Ilmn_1706531	0.38	0.10	$1x10^{-4}$	33.8	6.9	$1x10^{-6}$	1.67	0.44	$2x10^{-4}$
	ABCC5	Ilmn_1651964	0.30	0.14	0.03	30.7	9.7	$1x10^{-3}$	1.82	0.73	0.01
	ABCC5	Ilmn_2302358	0.25	0.31	0.40	-5.9	20.1	0.77	2.25	1.67	0.18
B	*PARL*	Ilmn_2257665	-	-	-	-1.3	8.0	0.88	-	-	-
	ABCC5	Ilmn_1706531	0.37	0.12	$2x10^{-3}$	30.3	8.7	$5x10^{-4}$	1.34	0.57	0.02
	ABCC5	Ilmn_1651964	0.03	0.16	0.84	7.8	11.4	0.50	0.46	0.91	0.61

Table 4.6. Phenotypic association with subcutaneous adipose *PARL*/*ABCC5* gene expression.
Three phenotypes, fasting IGR, visceral fat and T2D, regressed upon *PARL* and *ABCC5* transcript levels for the TwinsUK sample. Mixed linear models were used including a family identifier random effects intercept term and fixed effect terms for confounding variables age and experiment batch effects (3 level factor). Only subcutaneous adipose probe Ilmn_1706531 was associated with fasting IGR, visceral fat and T2D in univariate **(A)** and multiple regression analyses **(B)**. Transcript analysis sample sizes for IGR, visceral fat and T2D were 680, 619 and 820 respectively. Fasting IGR is quantile normalised with the β regression interpretable as standard deviation units. Visceral fat units are cm^2, while risk of T2D is on the logit scale, with $\exp^{(1.34)}$ equal to an OR of 3.8 (95% CI: 1.25 - 11.6).

Discussion

In this study, the non-synonymous *PARL* SNP rs3732581 (exon 7) and the *PARL* gene itself, showed no evidence of association with fasting plasma insulin and glucose levels or T2D for one African American and two European samples. By contrast however, the neighbouring *ABCC5* gene shows replicated evidence of genetic association with T2D and glycaemic traits for the same genetic location in multiple human populations, with the putative common functional variant in *ABCC5* playing a regulatory role in controlling *ABCC5* expression levels. These results are consistent with the idea that elevated *ABCC5* expression levels increase risk of insulin resistance, dyslipidemia and T2D (see Figure 4.3), since the same functional variant appears to be associated with not only *ABCC5* transcription levels, but also fasting IGR, visceral fat accumulation and T2D. In turn, *ABCC5* transcript levels are strongly associated with fasting plasma IGR, Homeostasis Model Assessment beta cell function and insulin peripheral sensitivity, visceral fat accumulation and T2D, which suggests that *ABCC5* expression is a causal risk factor for onset of T2D.

The *ABCC5* genetic variant association with T2D-related phenotypes may also explain previous contradictory human association results observed for Val262Leu, if observed association with *PARL* is in fact due to confounding LD with functional variants in *ABCC5*. For example, SNPs such as rs4912515 in immediate proximity to the *ABCC5* variant $\hat{S}$ location estimate show evidence of long range LD with Val262Leu in *PARL*.

Given the overwhelming evidence from animal model and functional studies for reduced *PARL* expression associated with observed and induced T2D (Civitarese *et al.*, 2010, Tang *et al.*, 2009, Walder *et al.*, 2005), how is this reconcilable with the results presented here for *ABCC5* and *PARL*? It seems that there is evidence for both genes to be to be implicated in T2D, but the key question for future research is to understand whether (and how) *PARL* and *ABCC5* are functionally related. At this point, the most likely explanation for these observations – due to the existence of a functional genetic variant associated with T2D in multiple human populations that regulates *ABCC5* expression – is that *ABCC5* has a causal (if modest) role in T2D onset, while *PARL* dysfunction appears to be a consequence rather than a cause of

insulin resistance and disease state. This is illustrated by *Psammmomys obesus* experimental studies (Walder *et al.*, 2005) - induction of T2D in these rats by high-fat diet and inactivity is associated with a reduction in *PARL* (skeletal muscle) expression, which can restored to normal levels by exercise training (running 1km/day in 1 hour for 3 weeks).

Empirically, *PARL* and *ABCC5* transcripts are correlated, suggesting co-regulation of the two genes due to shared transcription factors and/or *cis*-acting regulatory elements. It is also possible, but perhaps less likely, that *ABCC5* more directly regulates *PARL*. Based upon TwinsUK data, there is some indirect evidence of *cis*-co-regulation between the two genes with two *PARL* transcripts (in LCL) regulated by upstream eQTLs (located in *ABCC5* and *HTR3E*) and one *ABCC5* transcript (in skin) regulated by an eQTL located in *PARL* (or *PARL* promoter region). However, these observations are for tissues that are unlikely to be functionally relevant to T2D and therefore require verification. The results for adipose, which is the most relevant of the three tissues, only shows evidence for one *ABCC5* probe (Ilmn_1706531) regulated by an intragenic *cis*-eQTL.

ABCC5 is a member of the ATP-binding cassette transporter super-family (Dean and Allikmets, 2001), transporting cyclic nucleotides such as cGMP, which is a common regulator of ion channel conductance, glycogenolysis, and cellular apoptosis, making *ABCC5* a potential candidate for T2D. Glycogenolysis takes place in the cells of the muscle and liver tissues and is regulated by epinephrine, glucagon and insulin in response to blood sugar levels with hormone effects mediated by cAMP or cGMP (Lehninger, 1979). In myocytes, glycogen degradation serves to provide an immediate source of glucose-6-phosphate for glycolysis, to provide energy for muscle contraction; in hepatocytes, the main purpose of the breakdown of glycogen is for the release of glucose into the bloodstream for uptake by other cells.

Although ABC transporter genes including *ABCC5* have been well studied in relation to cancer treatment and drug resistance (Tamaki *et al.*, 2011), little is known about *ABCC5* in relation to T2D. One exception to this is an ABC transporter expression study (Nowicki *et al.*, 2008) of uptake and efflux transporters in the liver and kidney of streptozotocin-induced diabetic (SID) rats. The study demonstrates that the

quantity (Western blot) of Abcc5 (Mrp5) protein in the liver of SID rats fed a high fat diet is reduced to approximately 4% of the level in the control. In contrast, Abcc5 protein level in SID rats fed a normal diet were not significantly reduced compared to the controls. These observations indicate that the almost complete loss of the Abcc5 protein is in response to the combined effects of induced pancreatic beta cell failure and high lipid levels.

Furthermore, the study by Nowicki *et al.* (2008) adds support to further investigating the transporter gene family in relation to common T2D, with the sub-families of ABCC and ABCG already implicated in different ways (Matsuo, 2010). Mutations in the sulfonylurea receptor (SUR) 1 encoded by *ABCC8* have previously been shown to cause neonatal diabetes (Greeley *et al.*, 2011), MODY (Bowman *et al.*, 2012) and adult T2D (Tarasov *et al.*, 2008). The sulfonylurea receptor is a subunit of the ATP-sensitive potassium channels, which regulate insulin secretion from pancreatic beta cells by sensing cellular metabolic levels. In addition, the ABCG transporter genes have also been implicated in cholesterol transport in animal models of T2D (Levy *et al.*, 2010).

In summary, this study demonstrates that a common *ABCC5* functional variant observed in Europeans and African Americans is a risk factor for T2D. The mechanism for this association is likely to be that the genetic variant is a regulatory eQTL for *ABCC5* that confers risk of disease by altered *ABCC5* transcript expression levels. Increased *ABCC5* subcutaneous adipose expression levels are positively associated with fasting insulin levels, visceral fat accumulation and T2D. The association with intermediate phenotypes for T2D suggests that elevated *ABCC5* expression is a causal risk factor for T2D and not a consequence of disease onset. By contrast, *PARL* genetic variants and gene expression are not associated with intermediate phenotypes or T2D for human observational data, suggesting that previously observed and experimental evidence for association between *PARL* dysfunction and T2D is likely to be a consequence of disease onset. Given the *ABCC5* risk variant is cosmopolitan, common and observed in disparate populations of European and African ancestry, this indicates the polymorphism is very old and that the CDCV hypothesis is likely to apply to this gene.

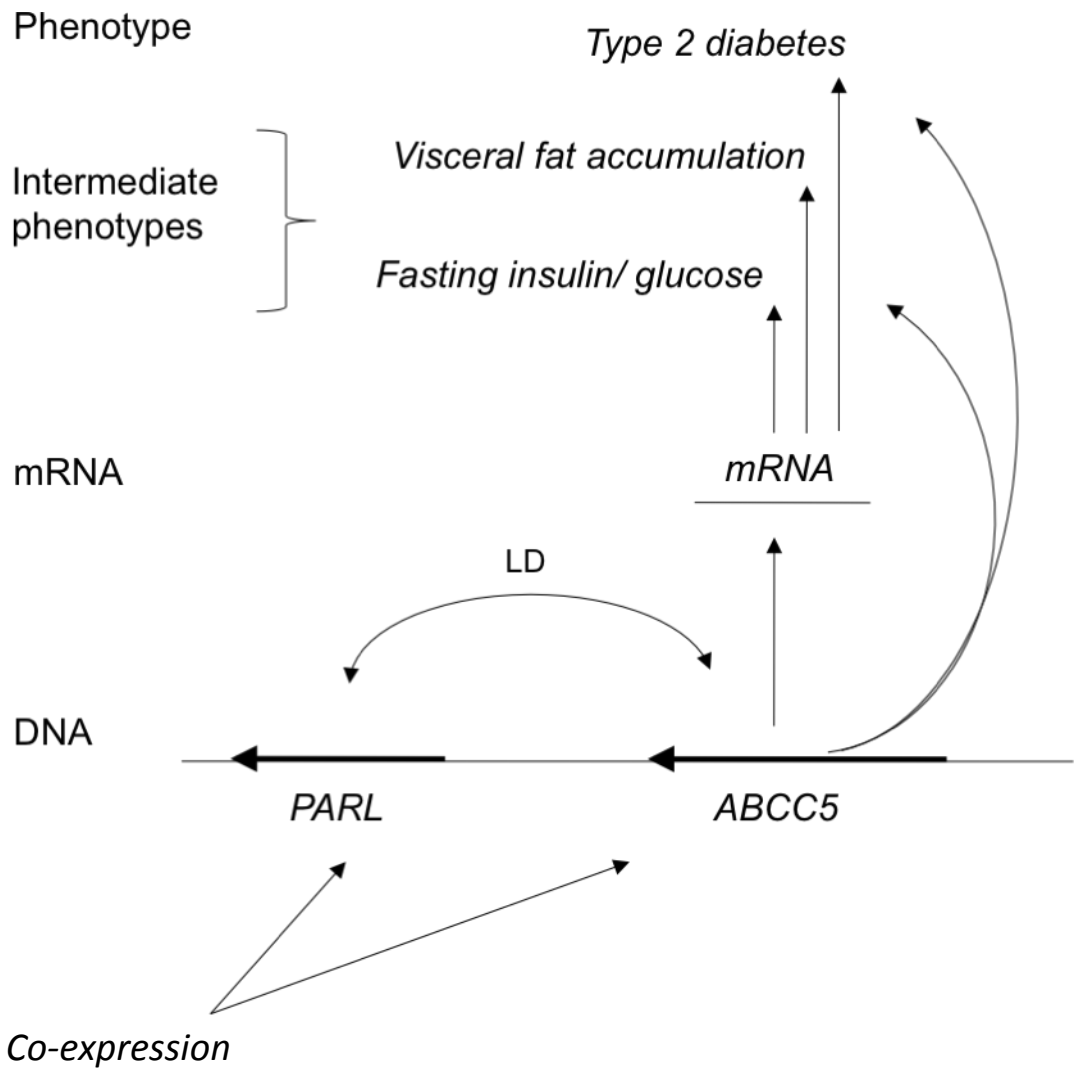


Figure 4.3. *ABCC5* genetic and mRNA association with intermediate phenotypes and T2D.

An eQTL in *ABCC5* (intron 26) regulates expression levels and is associated with fasting plasma insulin and glucose levels, visceral fat accumulation and T2D for TwinsUK data. *ABCC5* transcript Ilmn_1706531 is also strongly associated with intermediate phenotypes and T2D (Table 4.6). Consistent with the central dogma theory, this suggests the association between the *ABCC5* genetic variant is causally mediated via expression levels that are tissue and transcript specific. Double-headed arrows indicate correlation and curved lines association, but not directly causal.

Chapter 5: Mitochondrial Genetic Pathways and Susceptibility to T2D

Abstract

Changes to mitochondrial structure, function and quantity are routinely observed in tissues from type 2 diabetic subjects. Nuclear genes encode the vast majority of mitochondrial proteins yet just a handful of common variants within these genes are associated with T2D or related traits in GWA studies. Far more nuclear-encoded mitochondrial (NEM) genes may contribute to T2D but their effects may be too small to detect by testing each SNP individually. An alternative approach is to analyse groups of functionally related genes that form biological pathways. Pathway analyses have shown oxidative phosphorylation, which takes place within the mitochondrion, is down-regulated in T2D, however, several SNP-based pathway analyses failed to detect any association between mitochondrial pathways and T2D.

The study presented here attempts to identify NEM biological pathways associated with T2D. A total of 38 manually curated NEM gene sets were identified from public databases. Summary results from multi-marker, GWA scans for T2D in European and African American samples were then used to test for association with these gene sets. Five of the 38 gene sets tested show potential evidence of association with T2D in both European and African American data sets. T2D-associated NEM gene sets include those that represent pyruvate metabolism, the tricarboxylic acid cycle, and mitochondrial protein translation, collectively emphasising the role of ATP production in T2D aetiology.

Introduction

In this Chapter I present a pathway analysis of NEM genes to test for potential association with T2D. The rationale for a pathway analysis is that by simultaneously analysing a set of genes that represent a well-defined biological pathway, it may be possible to detect individually weak, but functionally coordinated changes in the biological processes that underlie complex traits. Several pathway-based studies in relation to T2D have been conducted (Perry *et al.*, 2009, Segre *et al.*, 2010, Mootha *et al.*, 2003). Here I present a modified pathway analysis based upon summary results from a powerful genome-wide analysis of European and African American samples using a multi-marker test of association.

Identification of NEM Genes

The first stages of the analysis are to define what constitutes a NEM gene and comprehensively identify them all. Multiple strategies have been employed to catalogue all NEM genes in the genome, these strategies can be broadly split into three types of information: sequence-based, mass spectrometry, and microscopy; for an excellent review of current methods, see Calvo and Mootha (2010). In this section I will provide an overview of these methods as it is important to understand for this Chapter.

One mechanism of importing proteins encoded by NEM genes from the cytoplasm into the mitochondrion is through a cleavable targeting amino acid sequence at the N terminal of the protein (Figure 5.1). These "post codes" are less than 100 amino acids (Neupert and Herrmann, 2007, Calvo and Mootha, 2010), and are read by transmembrane mitochondrial proteins that mediate the translocation; once within their pre-determined location the targeting sequence is cleaved from the protein.

Conservation of the targeting sequence in mitochondrial proteins allows identification of these proteins using bioinformatic tools, e.g. Target P (Emanuelsson *et al.*, 2000), MitoPred (Guda *et al.*, 2004), and MitoProt2 (Claros and Vincens, 1996) that interrogate amino acids sequence of nuclear proteins for conserved motifs. However, not all proteins are imported into the mitochondrion in this way and many non-mitochondrial proteins posses the mitochondrial-targeting sequencing. For these

reasons, algorithms based purely on target sequence alignment, i.e. without prior experimental evidence of mitochondrial localisation can suffer from high false positive and negative rates.

Mass spectrometry is another extremely fruitful technique to identify proteins. The first stage of this approach applied to the mitochondrial proteome is the isolation and purification of mitochondria, and then the population of proteins within this mitochondria-enriched fraction are analysed. Depending on the type of mass spectrometry being performed the proteins may or may not be digested with trypsin. Computational analysis of the resulting mass-charge ratio spectra can then be used to identify the proteins.

One of important analytical challenge for cataloguing the mitochondrial proteome through mass spectrometry is the lack of sensitivity (i.e. the probability of correctly detecting a mitochondrial protein, if it is present in the sample). Pagliarini *et al.* (2008) present data on the sensitivity of different analytical techniques based solely mass spectrometry in comparison to an integrative approach. Average sensitivity amongst the single methods compared was approximately 30%, which is interpretable as these methods detecting only 30% of mitochondrial proteome. Tissue specific expression of NEM gene and detection bias of the most abundant proteins in the sample may explain why this figure is so low.

A third approach to identify proteins localised in the mitochondrion is immunofluorescence microscopy, whereby proteins are visualised by the attachment of an immunoflorescent antibody. For large-scale purposes, such as at the proteome level, this method is laborious and results are potentially influenced by the antibody used. Hence, this method is more suited to relatively small-scale, focused studies.

In view of the technical and analytical challenges of any single method to catalogue the mitochondrial proteome, Calvo *et al.* (2006) have developed a whole-genome integrative approach to predict human mitochondrial protein localisation. The principle of this method is the creation of an aggregate score, termed “Maestro”, designed to reflect the probability that a protein localises to mitochondria. The score is calculated by combining information across eight sources of mitochondrial

evidence: presence of target sequences, protein domains, regulatory elements, yeast homology, ancestry, co-expression, PGC1 alpha-biogenesis induction, and mass spectrometry. Proteins with a Maestro score greater than a predefined threshold are considered to localise to the mitochondrion, with the threshold selected to maximise both sensitivity and specificity.

Pagliarini *et al.* (2008) extended the work of Calvo *et al.* (2006) by integrating mass-spectrometry and GFP tagging of the mitochondria proteome from 14 mouse tissues with the other prediction methods of mitochondrial localisation listed above (excluding biogenesis induction) to calculate Maestro scores. In total this compendium of mitochondrially-localised proteins, named MitoCarta, currently lists 1,098 mouse genes and their 1,013 human orthologs (http://www.broadinstitute.org/pubs/MitoCarta).

The data made available by Pagliarini *et al.* (2008), in addition to providing a comprehensive list of mammalian mitochondrial proteins, present fascinating insights into the differential expression of these proteins across various tissues. Of particular interest are that a third of genes encoding mitochondrial proteins are ubiquitously expressed, the number of mitochondrial proteins per tissue detected by tandem mass spectrometry range from 554 to 797, and mitochondrial content, as assessed by cytochrome C level, varies by 30 fold between tissues.

Table 1 | **Mitochondrial targeting and sorting signals**

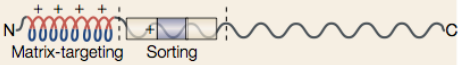
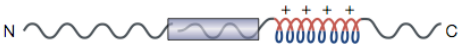
Signals	Preproteins	Import machinery	Location of signals in preproteins
Presequences			
At amino terminus of preproteins Cleavable upon import Enriched in positively charged, hydroxylated and hydrophobic amino-acid residues Amphipathic α-helix	Matrix proteins	TOM complex TIM23 complex	
Variations on presequences for sorting to mitochondrial subcompartments			
Non-cleavable 'presequence' plus hydrophobic anchor	Some outer membrane proteins	TOM complex	
Presequence plus hydrophobic anchor	Some inner membrane proteins	TOM complex TIM23 complex	
Bipartite presequence with dual targeting and sorting information: matrix-targeting sequence followed by more hydrophobic sorting signal (prokaryotic type)	Some proteins of inner membrane or intermembrane space	TOM complex TIM23 complex	
Presequence-like signal (positively charged) at internal position, often preceded by hydrophobic segment	Some inner membrane proteins	TOM complex TIM23 complex	
Multiple internal targeting signals			
Multiple internal signals in non-cleaved preprotein Charged and uncharged signals (unknown recognition motif) Signals can function independently but cooperate for efficient targeting and translocation	Metabolite carriers of inner membrane	TOM complex TIM22 complex	

Figure 5.1. Proteins are targeted to locations within the mitochondrion through encoded signals at the N terminus (Pfanner and Geissler, 2001).

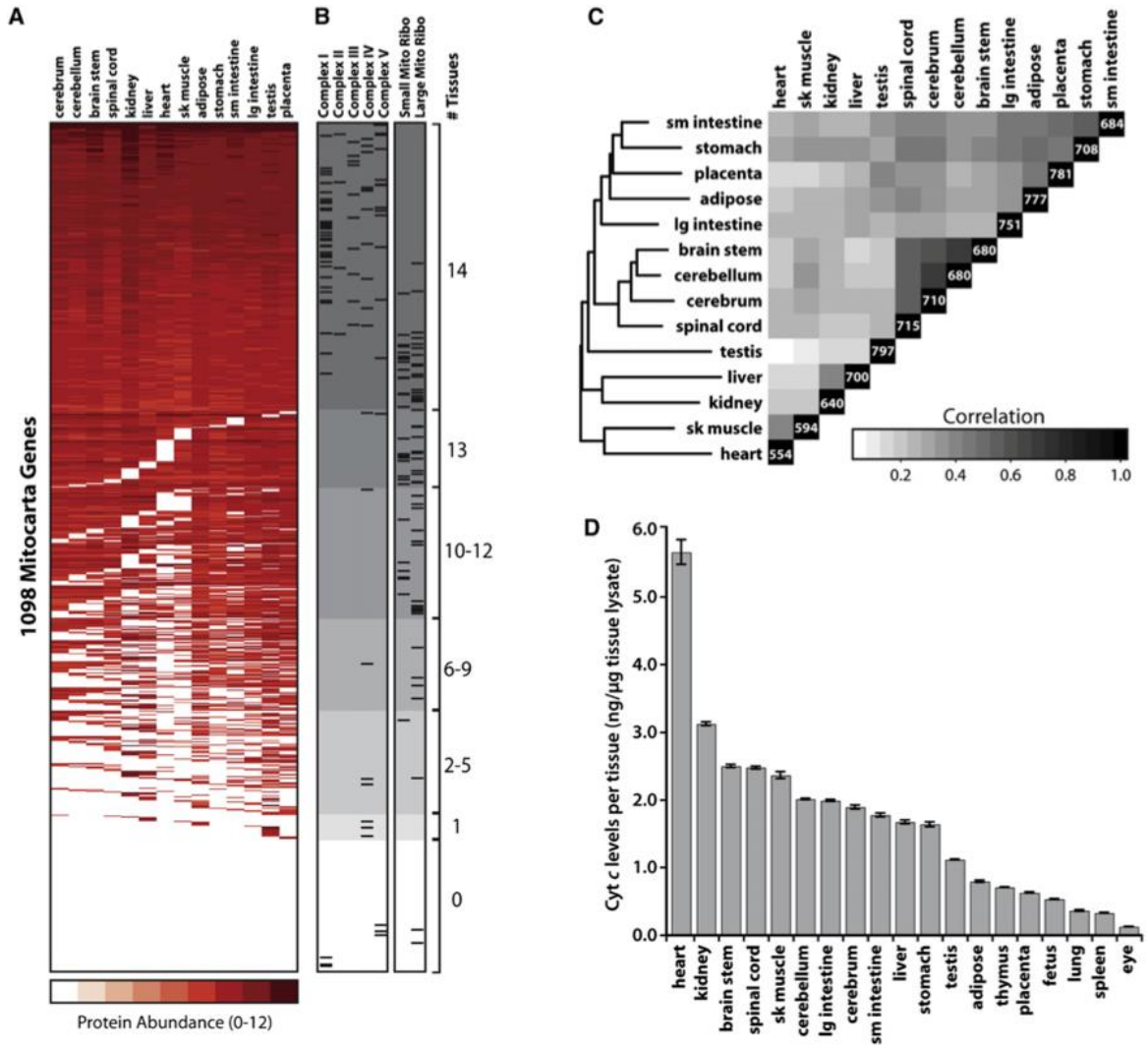


Figure 5.2. MitoCarta protein expression and mitochondrial quantity across 14 tissues (Pagliarini *et al.*, 2008).

Images A - D show potential tissue specificity by the nuclear-encoded mitochondrial genes (NEM genes) of MitoCarta. Approximately a third of all proteins encoded by NEM genes are ubiquitously expressed as detected by tandem mass spectrometry, whilst a quarter of the proteins encoded NEM genes could not be detected using this method (A-B). From the correlation matrix (C) it appears some of the tissues relevant to glucose homeostasis (adipose, liver, and skeletal muscle) show partial but not complete correlation of expressed NEM genes. The heart is by far the most mitochondrially-active tissue from the 14 considered here (D).

Pathway-based Analyses

The use of single marker association tests is widespread in dissecting the genetic architecture of complex diseases. With a few exceptions, the actual susceptibility variant(s) and their contribution to the trait, remain unidentified. Interaction between susceptibility variants within the same gene or between genes in biological pathways further adds to the difficulty in elucidating the biological impact of these variants. Therefore, many have adopted a more holistic approach of knowledge-based biological pathways to search for novel susceptibility loci. The set of genes that comprise a biological pathway can be used as framework to test for enrichment of association signal.

The basic scheme for pathway-based analyses of GWA data consists of five stages (Wang *et al.*, 2011). In the first stage, SNPs are mapped to genes by their transcriptional start and end coordinates. The start coordinate can be extended to the promoter region of the gene, the size of this region varies widely but a generic 1 Kb upstream of the transcriptional start site of the gene can be used to cover proximal and core elements of the promoter (Maston *et al.*, 2006). Regulatory motifs, such as enhancer, silencer, and insulator elements can be located much further away and at either end of the gene. Therefore deciding where a gene start and end depends on the context of the study, usually 5-10 Kb both upstream and downstream of the coding region is defined as a gene unit, but can be up to 500 Kb in both directions for regulatory regions of very large genes.

Each gene unit contains numerous SNPs each with a single association test statistic; pathway analyses require every gene to be represented by a single test statistic. Therefore, the multiple SNP associations for each gene must be collapsed into a single measure of association; the most frequently used method in the literature is to assign the most significant SNP p-value to the gene.

Once SNPs are mapped to genes and a suitable summary gene-level statistic is chosen, the second stage in a pathway analysis is the assignment of genes into molecular pathways. A plethora of public pathway annotation databases are available to facilitate this process, Table 5.1 list various databases as an example.

Selecting the most appropriate database depends on the research question and the available data, e.g. SNP, expression data, or methylation probe identifiers. Most, if not all, databases have cross-platform utility to convert standard commercial identifiers into a unique, unambiguous, internal identifier. Choice of internal identifier used in the database has implications for the conversion. For example, if multiple transcript probe identifiers can only be mapped to a single Ensembl or UCSC gene identifier, important information is lost through this form of reductionism. The transcript probes may discriminate alternative splice variants of the gene, some of which may have distinct or opposing roles.

Curation of biological pathways in these databases is of paramount importance in pathway analyses; curation can either be manual or automated. The latter far outnumbers the former in number of pathways deposited in databases and the automated pipeline has the potential for rapid incorporation of new data as it appears in the literature. However, this dependence on text mining algorithms may give rise to inaccurate pathways with a lack of human review and it is for this reason manually curated pathways are usually preferred where possible. Manual pathway curation is the result of human review of experimental work and thus regarded as the more reliable of the two.

Obtaining large-scale data is no longer a challenge in the field; the constraint lies in how the results are interpreted and only a proportion of biological pathways have been fully described (i.e. all components of the pathway characterised by functional work). This is the ultimate goal of pathway annotation. Many more pathways are understood to be incomplete and only more experimental work will allow their annotation to fully emulate the biological pathway. Until then this limitation is common to both manually and automatically curated pathways.

Name	Curation	Major features	URL
Biocarta	M	Driven by user input with expert review of some pathways	http://www.biocarta.com/
DAVID	M/E	Augments and integrates annotations from other databases	http://david.abcc.ncifcrf.gov/
GO	M/E	Largest database; hierarchical structure; can filter data by evidence codes	http://www.geneontology.org/
Ingenuity	M/E	Large collection of canonical pathways; high-quality pathway maps	http://www.ingenuity.com/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	M	Reference pathways (mosaics from several organisms) and organism-specific annotations; pathway maps link to closely related genes	http://www.genome.jp/kegg/
MetaCore	M	Extensive disease pathways; can edit pathway maps for publication	http://www.genego.com/
MetaCyc	M	Metabolic pathways; can visualize connections among pathways	http://metacyc.org/
Molecular Signatures Database (MSigDB)	M/E	Can download pathways from several other databases as a collection for input to analytical software; novel groupings (e.g. motif gene sets)	http://www.broadinstitute.org/gsea/msigdb/index.jsp/
PANTHER	M	Can predict protein functions from sequence and evolutionary data	http://www.pantherdb.org/
Pathway Interaction Database (PID)	M/E	Broad range of cellular pathways with special focus on cancer signaling; can generate interaction maps from a list of genes	http://pid.nci.nih.gov/
Reactome	M	Pathways are extensively cross-referenced to PubMed, HapMap and other resources; can overlay expression or other data onto pathway maps	http://www.reactome.org/ReactomeGWT/entrypoint.html/
ResNet Series	M/E	Regular updates through web server; optional user editing or text scanning of user documents; links to reference articles	http://www.ariadnegenomics.com/

Table 5.1. Public databases for listing molecular pathways and providing pathway analysis algorithms (Ramanan *et al.*, 2012)
Abbreviations: M – manually, E – electronic.

Allocation of genes to pathways in the databases is, in some aspects, an artefact of the studies and analyses performed and therefore may not recapitulate actual biological complexity. Each member of a pathway in database is given equal importance, which may not be true for all pathways; moreover there may be complementary mechanisms that rescue a pathway under certain conditions. In spite of this, manually curated pathways, as units of analysis, are our best attempt to reflect the biology.

Some of the pathway resources listed in Table 5.1 are briefly described here to provide an overview and background to the following section on statistical analysis of pathways.

Consistent gene descriptors across multiple databases are vital to study any biological pathway. The Gene Ontology (GO) project (Gene Ontology, 2013) is a major resource cataloguing standardised vocabulary to describe gene products. Gene products are described by annotation terms across three ontology domains: cellular component, molecular function, and biological process.

The hierarchical system of GO allows the user to define the level of information required. For example, if one were interested in glucose metabolism (GO:0006006) in humans there are currently 390 gene products under that GO term. Refining the search specifically to gluconeogenesis reduces the number of gene products to 85; further refinement of the GO search term to regulation of gluconeogenesis (GO:0006111) contains 24 genes. With more general GO terms, the number of gene products increases, as does the difficulty in determining the biological significance of the pathway analysis.

The quality of the annotation of a gene product can be assessed by its evidence code, which expresses how the annotation information was acquired, i.e. through experimental or computational methods, or in a very small proportion of annotations such information is not available (Rhee *et al.*, 2008). As of November 2012, the number of human gene products annotated was 44,914. A number of GO-based tools are available from the resource (Khatri and Draghici, 2005) that implement various statistical models.

Biocarta is an example of an exclusively manual curated database, where users can add new pathways or amend existing pathways after review. It is ultimately a database of biological pathways with no facilities to perform statistical analysis; therefore, its primary use is the retrieval of pathway information. Biocarta pathways are displayed as an interactive image whereby individual elements of the pathway are linked to additional information and other bioinformatics repositories. The list of genes in a pathway can also be viewed as text presented as full gene names with NCBI gene identifier, which is useful for gene set analysis.

The Kyoto Encyclopedia of Genes and Genomes, KEGG (Kanehisa and Goto, 2000), resource aims to provide a representation of biological systems (at time of writing, n = 272,804; http://www.kegg.jp/kegg/docs/statistics.html) across a range of organisms, despite its manual curation. KEGG presents this information across 15 databases spread across 3 broad categories of systems, genomic, and chemical information. Like BioCarta, KEGG also uses interactive pathway diagrams and provides additional pathway information.

The Molecular Signature Database (MSigDB) was developed as a resource for use in the gene-set enrichment analysis (GSEA), such as notable studies by Mootha *et al* (2003) and later by Subramanian *et al.* (2005). It is currently in its third major release (v3.1; October 2012) listing 8,513 gene sets across six collections reflecting their source – chromosomal location, manually curated, shared regulatory motifs, computational, gene ontological, and oncogenic signatures. Distribution of gene sets by collection in v3.1 is presented in Figure 5.3. The investigative gene set feature of MSigDB can be used to integrate a user's list of genes with biological pathways, where those genes appear most frequently and provide an informative summary of the extent of gene overlap between gene sets.

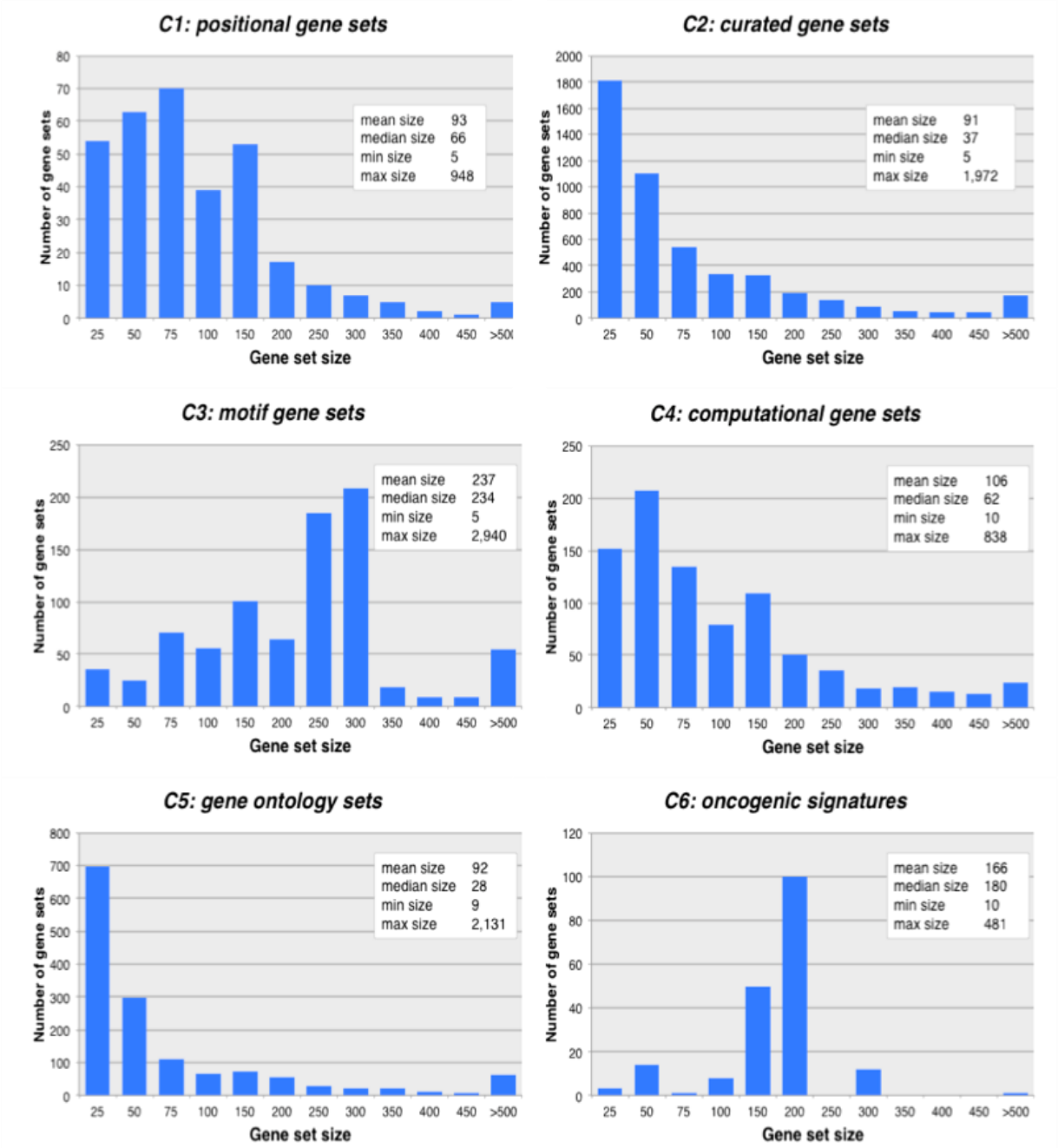


Figure 5.3. Frequency histograms of gene set sizes for each collection in the MSigDB.
Source: www.broadinstitute.org/cancer/software/gsea/wiki/index.php/MSigDB_Statistics

Gene identifiers from a variety of platforms can be ambiguous, resulting in multiple identifiers for the same gene or the same identifier corresponding to multiple genes. Like many gene set databases, MSigDB efficiently handles this potential ambiguity by mapping input gene identifiers to a common identifier. Alternatively gene identifiers can be converted prior to searching for gene sets using freely available tools such as BioMart ID converter, DAVID Gene ID Conversion Tool (Huang da *et al.*, 2008), and IDconverter (Alibes *et al.*, 2007) to name but a few.

Internal identifiers need to possess certain properties that facilitate their use, for example, they need to be unique, unambiguous - i.e. provide a 1:1 correspondence between identifier and gene, stable, and well integrated within many biological databases. The MSigDB uses the Entrez gene identifier as the internal identifier because it fulfills all these requirements.

MSigDB was chosen for this study for a number of reasons. The first being it is freely available to the public, secondly because it allows the user to batch query a list of genes and automatically search for overlap with gene sets within the database. Thirdly, the MSigDB contains the largest number of manually curated gene sets (at the time of writing, 4,850 gene sets), which is considered the most reliable form of curation. The latter point is also related to the high-level of integration with other biological databases, including BioCarta, KEGG, Reactome, and the Pathway Interaction Database.

Hypothesis

The third stage of a pathway analysis is to decide on what hypothesis is being tested. Goeman and Buhlmann (2007) describe two types of pathway analysis - competitive and self-contained tests, differing in their definition of the null hypothesis and interpretation. In a competitive test, the null hypothesis is that the association of genes of a particular pathway are no different from the association of genes selected at random from the genome (matched for size).

A self-contained test simply asks whether there is an association between a pathway and the trait, with the null hypothesis being no association. A pathway p-value is

constructed by combining the p-values for all genes in the pathway, for example by Fisher's method, and the pathway p-value is then compared to a standard probability distribution. The test is described as self-contained as the input data is only from genes of a single pathway; association information from genes outside that pathway is not required.

Statistical Analysis of Pathways

Stage four of a pathway analysis is the statistical test. Traditionally pathway analyses were designed to be applied to genome-wide expression microarray experiments. For example, translating a list of differential expressed genes between two conditions (e.g. case-control, exposure to a drug, etc.) into appropriate biological pathways. These pathway methods have been adapted to analyse data GWA scans.

The simplest form of pathway analysis compares a list of genes deemed significant with known pathways. The null hypothesis being tested is that the list of genes of interest (e.g. significant SNPs from a GWA scan) is not significantly overrepresented in the pathways tested. In practice it is a count of genes on the significant list in a pathway compared with a count of the list not on the pathway that are formally compared using a 2 by 2 table contingency table (Table 5.2) and a p value can be obtained by either the Fisher's exact or Chi-squared test.

	Genes associated with phenotype	Genes not associated with phenotype	
Genes in pathway	a	b	a+b
Genes not in pathway	c	d	c+d
	a+c	b+d	N

Table 5.2. Two by two contingency table - statistical measures that can be used to test for association between phenotypic and molecular pathways.

The test statistic, χ^2, is given by:

$$\chi^2 = n(ad - bc)^2 / (a+b)(c+d)(a+c)(b+d)$$

Which is equivalent to the Pearson's χ^2 but avoids calculating expected values:

$$\chi^2 = n(ad - bc)^2 / (a+b)(c+d)(a+c)(b+d)$$
$$= \Sigma(O_i - E_i)^2 / E_i^2$$

As the proportion overlap between a pathway and the list of significant genes (cell a, Table 5.2) is small under the null, the Fisher exact test is more appropriate. The probability of the obtaining the observed values under the null is calculated as follows:

Fisher's exact test:

$$\text{p-value} = (a+b)!(c+d)!(a+c)!(b+d)! / a!b!c!d!n!$$

An obvious question for a competitive test is, what list of genes is the pathway genes compared to? The comparison list of genes could be all genes of the genome or restricted to those that were actually tested for association. The latter option is the more rational approach as it avoids any possibility of introducing genes that were not included in the analysis. For example, if SNP data is used, the comparison list of genes should be confined to genes with SNPs that have passed quality control (QC), which is a preliminary step in all genome-wide studies, but is important to remember for these purposes. Genes without any SNPs that have passed the QC filter are should not be represented in the analysis.

A number of groups have explored gene-set based overrepresentation approaches with a variety of gene scores and test statistics, some of examples are listed in Table 5.3.

Study	**Gene Score**	**Gene Test Statistic**	**Significance Assessment**
Wang *et al.* (2007)	Most significant SNP p-value	Kolmogorov–Smirnov	Phenotype permutation
Torkamani *et al.* (2008)	Most significant SNP p-value	Fisher's exact test	Hypergeometric distribution
Askland *et al.* (2009)	Most significant SNP p-value	Fisher's exact test	Gene set permutation
Holmans *et al.* (2009)	Most significant SNP p-value corrected for gene size	Fisher's exact test	Gene set permutation
Peng *et al.* (2010)	Fisher's combination test	Fisher's exact test	Hypergeometric distribution
	Sime's combination test	False discovery rate	

Table 5.3. Defining features of selected overrepresentation approaches (Wang *et al.*, 2011).

Gene Set Enrichment

Alternative to having to specify a significance threshold, enrichment methods rank all genes in order of significance, and then test for differences between ranks of genes in a pathway compared to non-pathway genes. Mootha *et al.* (2003) introduced GSEA as part of a gene expression microarray study of T2D; the approach starts by ranking genes (one expression probe per gene) according to difference in gene expression between normal and disease states. A schematic diagram of this non-parametric statistical test is presented in Figure 5.4. The gene ranking is used to construct a running sum statistic for each gene set, from the first ranked gene, the test statistic increases or decreases by a value depending on whether the gene is a member of the gene set (S) being considered. The maximum enrichment score (MES) for each gene set is defined as the largest positive deviation of the running sum. Significance of the MES is assessed by permutation of phenotype labels and recalculation of the MES; the p-value is calculated as the fraction of MES from the permutated data sets that are greater than the observed MES. The process is repeated for all gene sets being considered.

Through GSEA, Mootha *et al.* (2003) identified the oxidative phosphorylation gene set as being down regulated in T2D. Further examination of the gene set revealed a subset that accounted for the majority of the MES for the oxidative phosphorylation gene set and that the expression of this subset was reduced in individuals with impaired glucose tolerance but not normoglycaemic. This study demonstrates the effectiveness of gene set methods over single gene methods in that conventional microarray analysis were unable to detect any genes that showed significant differential expression between normal and disease states, whereas GSEA identified an entire biological pathway that contributes to the disease using the same data.

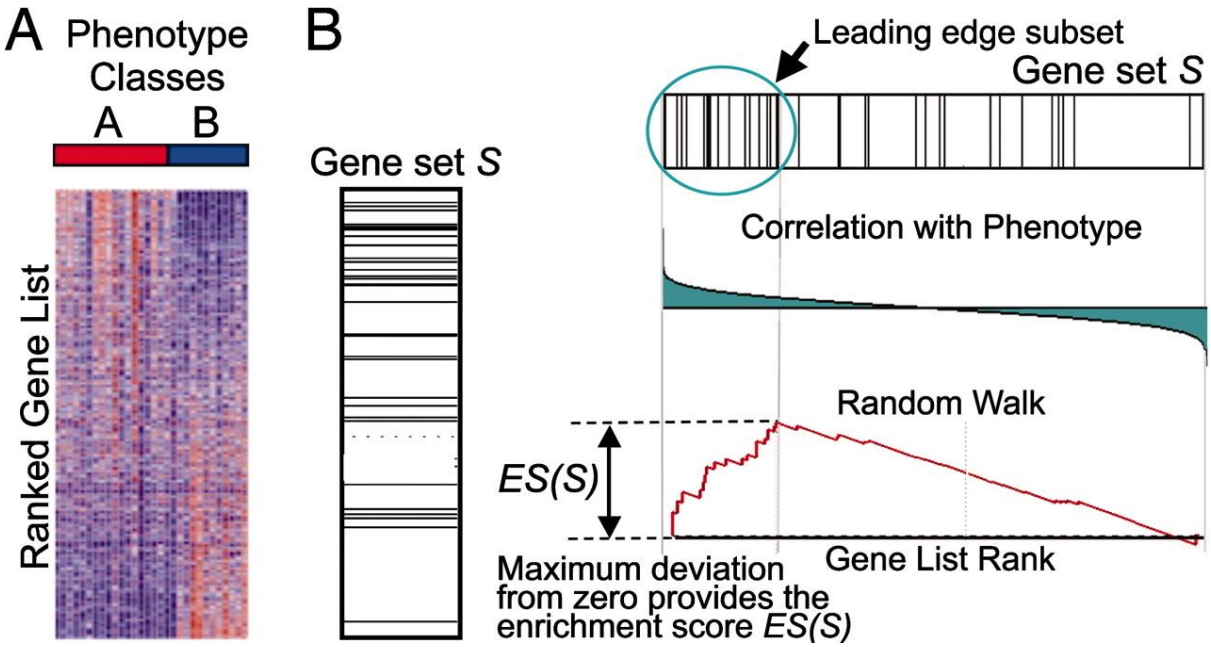


Figure 5.4. Gene set enrichment analysis (Subramanian *et al.*, 2005).
This figure outlines the method of gene set enrichment analysis (GSEA). The procedure begins with genome-wide, case-control microarray expression data; genes are ranked in accordance with the differential expression between the two phenotype classes (A). For a particular gene set, S, the aim is to determine if genes of S appear predominantly at the top or bottom of the ranked list, or are uniformly distributed. The enrichment score of the gene set, ES(S) is calculated by walking down the ranked list and increasing the running score whenever a gene from S is encountered and conversely it decreases when genes from S do not appear. The score is weighted by phenotypic correlation/association meaning that the score increases/decreases more largely for genes of S at the top and bottom of the ranked than those in the middle of the ranked list.

Bias

There are potential sources of bias in each of the five stages of a gene set analysis, some can go unnoticed, without any effect on the final results, whereas others have effects that cloud any putative signal. The allocation of SNPs to genes is relatively straightforward; one minor consideration is that of nested genes, that is the spatial arrangement of one gene within another gene. A gene nested within another maybe fully contained within an intron or spread across introns and exons of the larger gene. Based on a 2004 build of the human genome, Yu *et al.* (2005) estimated there were 373 nested genes in the human genome. A simple one gene nested with a larger host gene is considered here for simplicity, but other possibilities such as multiple nested genes with the same gene or more complicated scenario of two physically overlapping genes (Kumar, 2009).

If a SNP is assigned to both genes and both are biologically related, i.e. act in the same pathway; this will give rise to false evidence of enrichment for that particular pathway. A preliminary check in the analysis can reveal overlapping gene boundaries within each pathway under analysis. As nested genes makeup a very small proportion of all genes in the genome, they are the least important source of bias.

A second SNP-related bias is the variability of SNPs between genes. Genome-wide arrays are design to maximise coverage of SNPs with the minimum number of SNPs through LD. Because LD represents lack of recombination and recombination rates vary across the genome, genes of the same physical size will vary on a genetic scale and the number of SNPs required for adequate coverage. Genes with many SNPs will be expected to show a greater number of significant associations than genes with fewer SNPs – a correlation confirmed by Hong *et al.* (2009).

Local clusters of genotyped SNPs, such as those within a gene, may show a correlation structure, that is, they are not completely independent of each other. Genotyped SNPs in high LD between two or more genotyped SNPs is a form of redundancy and can be reduced by removing SNPs that are above a certain LD threshold. An alternative is to correct the gene test statistic for the number of non-independent tests being performed with an appropriate LD threshold

In GWA studies it is unlikely that a casual susceptibility locus will be represented on the array. Instead, it is more likely that a marker SNP is in LD with the putative susceptibility locus. The same logic applies to and is extended in gene-based analyses - in addition to the one or more genotyped SNPs tagging an non-genotyped variant, a variant driving the association signal may not necessarily reside within the gene, but could be located some distance away and be in LD with one more SNPs within that region.

Non-independence in gene set analyses also occurs within and between gene sets; members of the same gene set are likely to be non-independent in terms of expression. And when testing more multiple gene sets, some gene sets maybe more related to others through shared members. Non-independence between genes in a pathway being tested violates an assumption of most standard statistical tests, which usually assume identical, independent distributions (IID) for each data point. This includes hypergeometric tests and Fisher's exact test. A positive correlation between the genes of a pathway can artificially inflate the apparent association with a trait. Under the null hypothesis, the probability distribution is no longer hypergeometric (Goeman and Buhlmann, 2007) for correlated genes and the test produces an inflated p-value. Goeman and Buhlmann (2007) simulate how intra-pathway gene dependencies affect the p-value of a pathway; the proportion of pathways rejected under the null increases for correlations of 0.2 through to 1. For example, at a correlation (r) of 0.2 and alpha threshold of 0.05, the increase in proportion of pathways rejected under the Null compared to $r = 0$ is 1.9 fold; for the same r of 0.2 but an alpha of 0.0001, the increase in proportion rejected is 2.24 fold greater. The trend exhibits greatest effect for tests with smaller alpha threshold and large correlations.

Gatti *et al.* (2010) investigated the issue of intra pathway dependency using 202 data sets from the NCBI Gene Expression Omnibus data repository. They confirm the ubiquitous presence of inter-gene/intra-pathway correlation and find that violation of the independence assumption can increase the false positive rate up to 80%. The authors argue that in a expression microarray context, permutation of the gene labels handles intra-pathway correlation and the significance of the observed pathway test statistic is determined through comparison with an empirical null distribution of pathway test statistics.

One final important source of bias is size variation for genes and gene sets. The size of a gene can vary on both genetic and physical scales; for example, genes of similar physical length may vary widely on a genetic distance, which is related to the underlying rate of recombination in the regions where the genes are located. If an additive genetic map is available, gene sets can be compared in terms of total genetic and physical lengths. The total length of a gene set is a combination of the length of individual genes and the number of genes in the gene set.

Materials and Methods

Subjects and Data

This study analysed the GWA results from European (WTCCC1) and African American (National Institute for Diabetes and Digestive and Kidney disease) LD mapping studies for T2D (Lau *et al.*, in preparation). LD mapping has been described in detail elsewhere (Maniatis *et al.*, 2002, Zhang *et al.*, 2002, Maniatis *et al.*, 2004). The main advantages of this approach are (1) it does not assume that the susceptibility variant is genotyped nor in high LD with a genotyped marker; and (2) facilitates informative multi-marker tests of genetic association with a phenotype. This approach has recently been applied to Crohn's disease and identified more than 200 susceptibility loci (Elding *et al.*, 2011, Elding *et al.*, 2013), far surpassing the number from conventional GWA studies for Crohn's disease.

Definition of NEM Genes

The entire list of 1,024 human NEM genes was downloaded from MitoCarta (Pagliarini *et al.*, 2008), http://www.broadinstitute.org/pubs/MitoCarta/). Entries with duplicate Human Entrez identifiers were removed (n = 11), as were entries located on the mitochondrial genome (n = 13), X chromosome (n = 30) and Y chromosome (n = 1). This filter returned 968 unique Entrez identifiers. Gene coordinates provided by MitoCarta were UCSC human assembly March 2006 (hg18) and were converted to UCSC human assembly May 2004 (hg17) to be compatible the genomic data used. Build conversion was performed using the LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver); all 968 gene coordinates were successfully mapped to hg17.

NEM Gene Sets

Pathway membership for each NEM gene was identified through the MSigDB (http://www.broadinstitute.org/gsea/msigdb). Symbols for each known NEM gene were uploaded to the MSigDB: 955/968 submitted gene identifiers were recognised by the database and the top 100 overlapping gene sets with the manually curated collection (C2) were obtained. From this list of 100 gene sets, any gene set with less than 50% of NEM gene members were filtered out, resulting in 37 gene sets taken for analysis. This threshold of 50% was used to prioritise gene sets in which NEM genes

made up the majority of the biological pathway. In addition, the full MitoCarta NEM gene list was also included as a separate gene set, in total 38 gene sets were analysed. All genes of NEM gene sets were included in the analysis including non-NEM genes.

NEM Gene Set Tissue Specificity

There is evidence of tissue specific expression of NEM genes at the protein level (Pagliarini *et al.*, 2008). To potentially identify tissue-specific NEM gene sets related to T2D – namely, adipose, liver, and skeletal muscle – pathway gene memberships were obtained for each list of expressed NEM genes by using MSigDB as described above. As the tissue specific lists are subsets of the original NEM gene list, the majority of tissue specific NEM gene set memberships were the same as obtained for the full list of NEM genes; however, a novel gene set that met the definition of a NEM gene set (i.e. >= 50% of genes are NEM genes) in the liver sample of NEM genes was identified.

Genomic Regions

Using the location estimate for each analytical window, an additional 2 LDUs were added to each side of the functional variant estimate to create the genomic region of interest; the value of 2 LDU was chosen as values greater than 3 reflect large genetic distances with substantial breakdown in LD. LDU locations were interpolated onto a physical map so that genes could be allocated to genomic regions by overlapping start and end physical coordinates. Build 35 was used for the European data set and build 36 for the African American.

Gene Set Size

Gene set size is the number of genes of a gene set; not all genes of a gene set fall within the genomic regions as defined above, reducing the gene set size. Supplementary Tables 5.1 – 5.4 list all 38 NEM gene sets considered for analysis along with the gene set size, number of mapped genes and genes within genomic regions for the European and African American data sets.

The average percentage of genes from NEM gene sets that could be mapped to the genome was 92.20%; for a small number of cases (n = 4), this figure exceeded 100%,

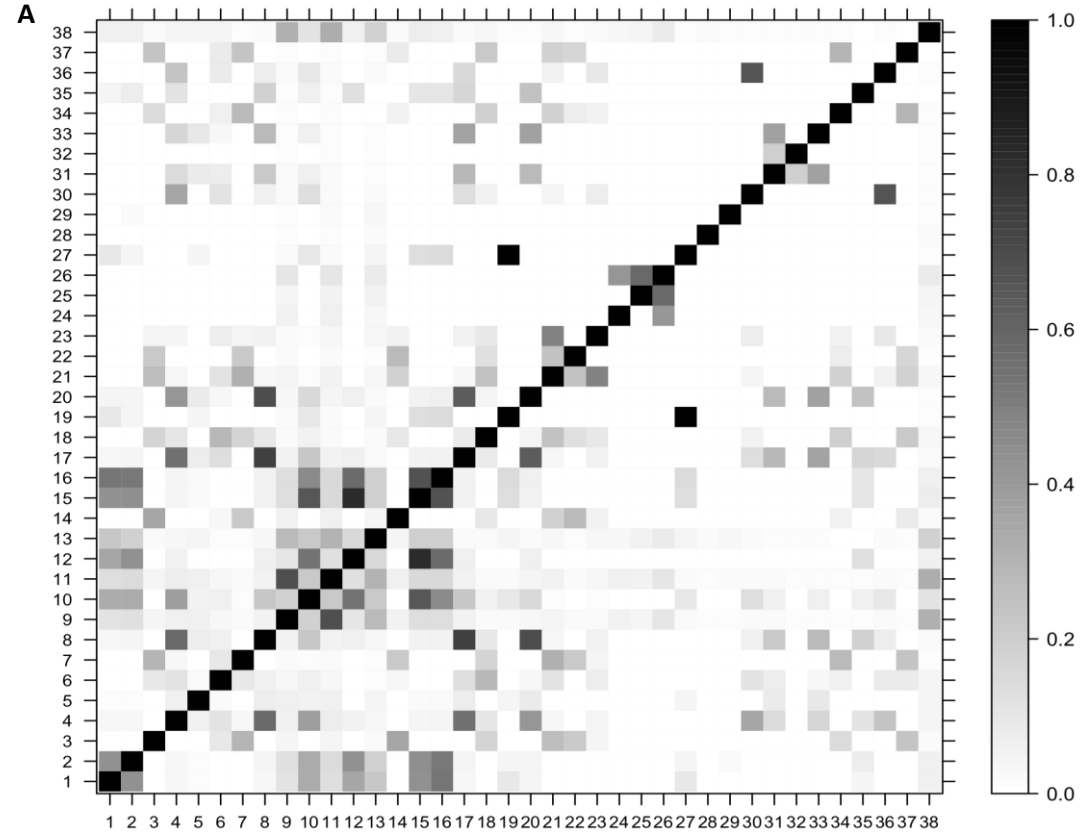
which is due to Ensembl returning multiple start and end coordinates for a single gene and likely to reflect the presence of more than one viable transcriptional start and stop site. As the differences between the alternative sites were very small, both sets of coordinates were kept.

Rather than number of genes of a NEM gene set, of greater importance for this analysis is the number of genomic regions that contain genes from a particular NEM gene set. I refer to this as the effective gene set size, because genes located outside the defined location estimate genomic region (but still within the analytical window) are unlikely to be responsible for any observed association with T2D. The Ringo package (Toedling *et al.*, 2007) was used in R (R Development Core Team, 2011) to compute overlap between genomic regions and gene coordinates. For each NEM gene set, a binary label was assigned to each genomic region to indicate the presence (1) or absence (0) of genes from the NEM gene set.

Where a gene overlaps neighbouring genomic regions, the gene is assigned to both genomic regions. There is an approximate 1:1 correspondence between the number of genomic regions with genes of a NEM gene set (effective gene set size) and the number of genes within the genomic regions (Supplementary Table 5.3). The average percentage of genes from NEM gene sets within genomic regions for the European and African American data sets were 34% and 25% respectively.

Extent of NEM Gene Set Overlap

To correctly address the issue of multiple testing using a Bonferroni correction, it is important to quantify how many independent tests are performed. Overlap between two NEM gene sets was calculated by combining the unique genes into one list and expressing the proportion of that list shared by both gene sets. The overlap results (Figure 5.5) were plotted using the Lattice package (Sarkar, 2008) in R. As the figure illustrates, there is very little overlap between the 38 NEM gene sets, which means the tests are essentially independent and therefore a Bonferroni correction is valid. The Bonferroni corrected alpha threshold at 5% for 36-38 independent tests is approximately $1x10^{-3}$ ($0.05 / 38 \approx 0.05 / 36 \approx 1x10^{-3}$).


A
38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
1.0 0.8 0.6 0.4 0.2 0.0

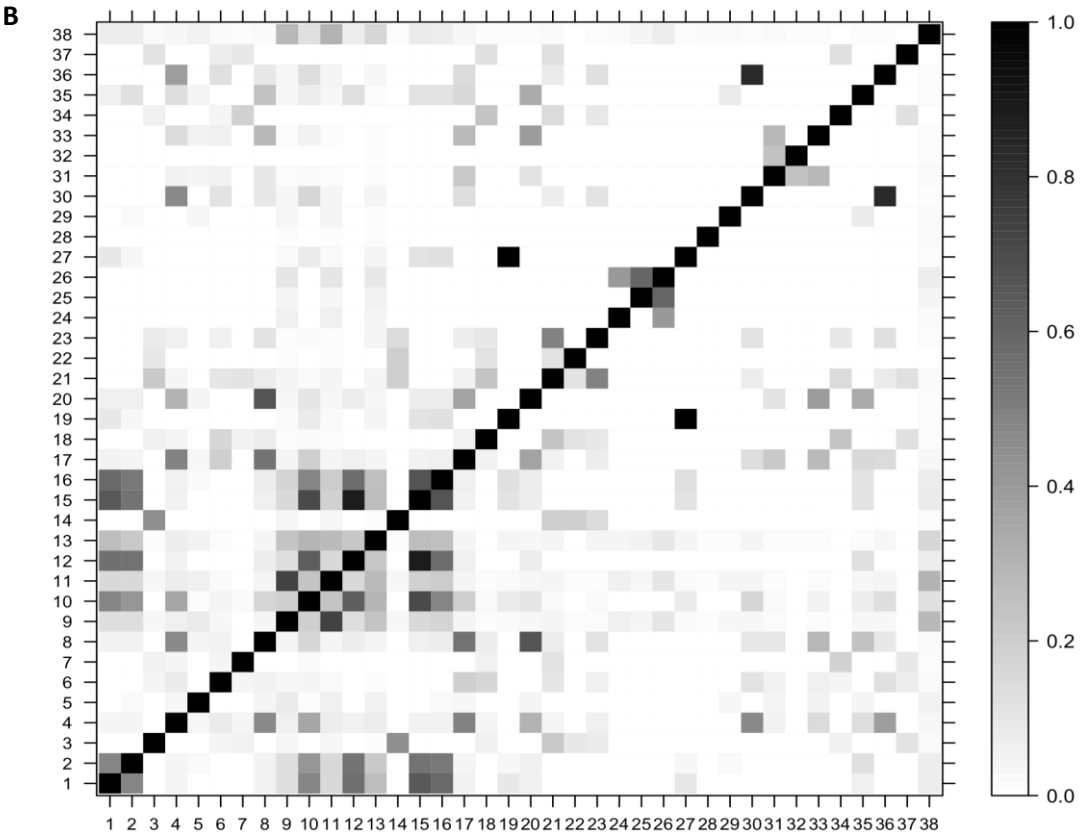

B
38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
1.0 0.8 0.6 0.4 0.2 0.0

Figure 5.5. NEM gene set overlap for European and African American data sets.
Plots A and B illustrate the extent of overlap between all 38 NEM gene sets. The overlap is expressed as the proportion of genes shared between any two NEM gene sets, 0 indicates that no genes are shared, whilst 1 indicates all genes are shared. The gene sets are numbered as listed in Supplementary Table 5.1)

Set-Based Analysis

For each of the 38 NEM gene sets, a two-sample t-test was used to test if the mean χ^2 association statistic for T2D was greater for genomic regions that contained genes of the NEM gene set against those that did not. An empirical p-value was obtained for each NEM gene set by permuting gene-membership status 100,000 times, the empirical p-value is given by the proportion of permuted test statistics that are equal to or more extreme than the observed test statistic: $\sum(t_{n..N} > t_0)/$ N. Where, $t_{n..N}$ represents the test statistic (t) from each permutation, t_0 is the observed test statistic, and N is the total number of permutations.

Over-Representation Analysis

For comparison with the quantitative analysis above, which compared the distribution of chi-square statistics for NEM gene set versus non-NEM gene set genomic regions, a more qualitative test was conducted using a 2x2 contingency classification. Four counts were observed in this table: (1) genomic regions that were nominally associated (p-value < 0.05) with T2D and contained genes of the NEM gene set, (2) genomic regions that were nominally associated with T2D and did not contain genes of the NEM gene set, (3) genomic regions that were not associated with T2D and contained genes of the NEM gene set, and (4) genomic regions that were not associated with T2D and did not contain genes of the NEM gene set. Fisher's exact test as implemented in R was used to test for association between presence of NEM gene set members and T2D. Fisher's exact test as opposed to the contingency χ^2 test was used because it was anticipated that some cell counts would be small (expected cell count < 5), which violates one of the assumptions of the χ^2 test.

Results

Descriptive statistics of the 38 NEM gene sets outlined in the Materials and Methods section of Chapter 5 can be found in Supplementary Tables 5.1 – 5.4. The Mean, median, and inter-quartile range of the effective gene set sizes for the European data set were 33, 13, and 21 genes respectively. For the same gene sets, the African American data set had a smaller effective gene set sizes due to the relationship between physical and genetic maps. This is observation is explained by the African population having shorter stretches of LD due to being much older with more historical recombination than the European population, which is the primary cause for breakdown in LD. As a result, when genetic distances are interpolated onto a physical map they are usually shorter for the African American population. Mean, median, and inter-quartile range for the African American data set were 36, 10, and 17 genes respectively. Two gene sets were removed from the African American data set due to mapped gene set size being less than or equal two genes, thus, only 36 of the 38 defined NEM gene sets were analysed for this data set.

Over half of the gene sets analysed (n = 21) showed association with T2D-implicated loci at an alpha threshold of < 0.10 in the European data set. Replication of these results was sought in an independent data set using African American T2D results; analysis in this data set produced 19 out of 36 nominally significant gene sets. A gene set was considered to be potentially replicated if it is present in both European and African American data sets with a p-value less than or equal to 0.10 or replicated, if the meta-analysed p-value < $1x10^{-3}$. A threshold p-value of 0.10 was used at this stage to reduce the probability of over-looking potential association between a gene set and T2D. There were 12 gene sets that satisfied this criterion. Analysing the same data using overrepresentation analysis in the form of Fisher's exact test on a 2x2 contingency table produced just three replicated gene sets below the same significance threshold.

To quantify the relationship between effective gene set size and p-value for these data (i.e. inflated association due to large gene set size), all manually curated gene sets deposited in the MSigDB (4,850 gene sets as of release 3.84, October 2012) were analysed. A multiple regression model that included both linear and quadratic terms

for effective gene set size explained approximately 37% of the variance in gene set – $\log_{10}$ p-value. A simpler linear model omitting the quadratic term had an adjusted R^2 of 35% (Table 5.4). Figure 5.6 illustrates that the regression line from the full model including a quadratic term fits the data only slightly better than the linear model gene set sizes greater than 300.

Model	Term	β	Standard Error	t statistic	p-value	Adj. R2
A	Intercept	4.81×10^{-3}	8.81×10^{-3}	54.64	$<2 \times 10^{-16}$	35.13
	Set size	5.97×10^{-3}	1.17×10^{-4}	50.81	$<2 \times 10^{-16}$	
B	Intercept	0.42	9.92×10^{-3}	42.20	$<2 \times 10^{-16}$	37.32
	Set size	8.77×10^{-3}	2.46×10^{-4}	35.69	$<2 \times 10^{-16}$	
	$(\text{Set size})^2$	-8.25×10^{-6}	6.38×10^{-7}	-12.94	$<2 \times 10^{-16}$	

Table 5.4. Linear and quadratic regression models for the effect of gene set size on gene set p-value.
In both the simple linear regression (A) and the quadratic model (B), effective gene set size has a highly significant effect on the p-value of the gene set and explained to 35-37% of the variance in the gene set p-values. These regression coefficients were applied to the 38 NEM gene sets and gave identical results in terms of NEM gene sets that were significantly associated with T2D in European and African American data sets.

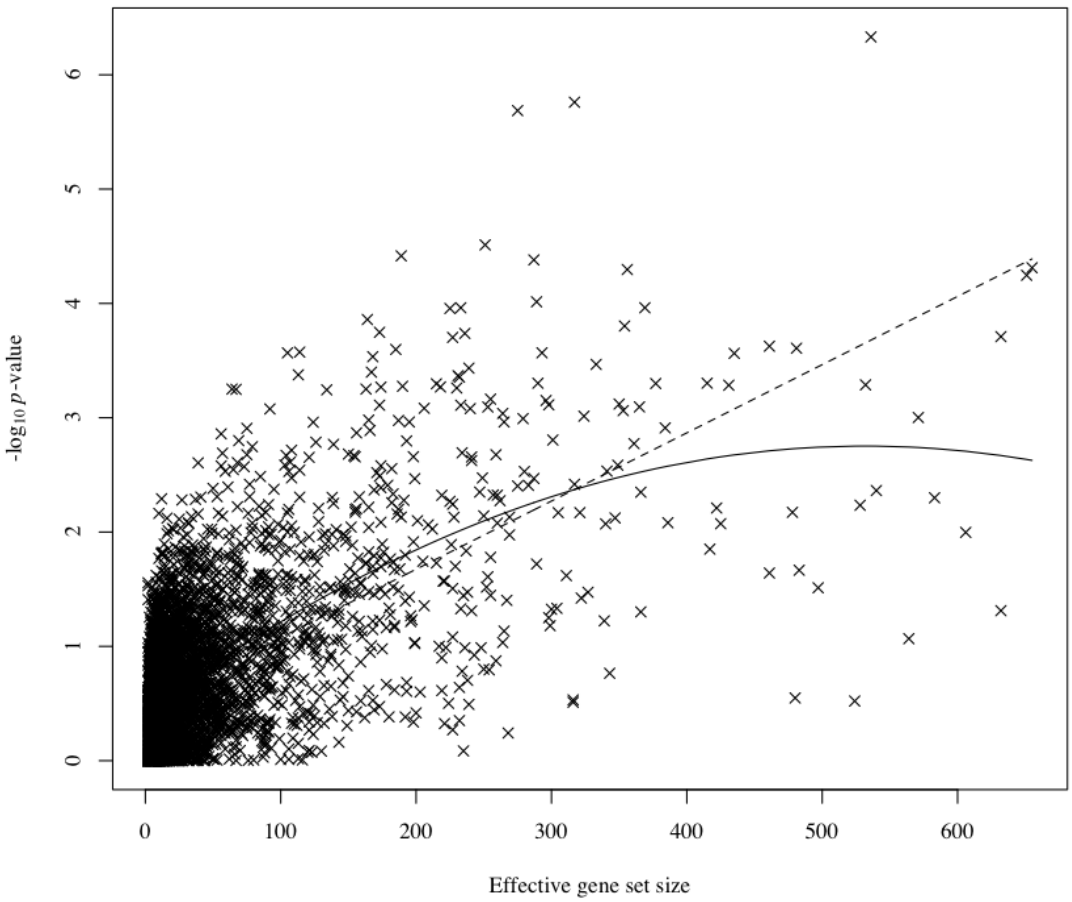


Figure 5.6. A plot of all manually curated MSigDB gene sets (n = 4,850) association with T2D (-log10 p-value, y axis) on effective gene set size (x axis).
The dashed regression line represents the linear regression model omitting the quadratic term for effective set size, whereas the solid line represents the regression model with both linear and quadratic terms for gene set size.

On the basis of model-fit statistics, the quadratic regression model was a marginally better fit than the linear model. Alternative effective gene set size correction factors were calculated based on the both linear and quadratic models and compared. The effective set size corrected p-value was calculated by subtracting the fitted $-\log_{10}$ p-value from the model from the observed $-\log_{10}$ p-value. The two correction factors gave identical results in terms of NEM gene sets considered significantly associated with T2D in the individual data sets and replication. For the European data set, 15 NEM gene sets were significantly associated with T2D after correcting for effective set size; a similar result was found in the African American data set, whereby 12 NEM gene sets were significantly associated with T2D. Five NEM gene sets corrected for effective set size were significantly associated at an alpha of 0.10 in both data sets (Table 5.5).

Fisher's method (Fisher, 1925) was used to meta-analyse p-values from these five NEM gene sets and these are presented in Table 5.5. As 36-38 independent NEM gene set tests were performed, the appropriate significance threshold is an alpha of $1x10^{-3}$, at which only one NEM gene set (NEMG988) is considered significantly associated with T2D.

Gene Set	European				African American				Meta analysis
	Effective Set size	t statistic	p-value	Adj. p-value	Effective Set size	t statistic	p-value	Adj. p-value	
NEMG988	324	3.29	1.00×10^{-5}	2.49×10^{-3}	248	2.85	2.50×10^{-4}	3.06×10^{-2}	0.0008
Wong Mitochondria Gene Module	91	2.57	5.00×10^{-4}	7.04×10^{-3}	56	1.84	1.00×10^{-2}	7.67×10^{-2}	0.005
Reactome Pyruvate Metabolism and Citric Acid TCA Cycle	17	1.95	3.08×10^{-3}	1.13×10^{-2}	12	1.40	2.88×10^{-2}	9.61×10^{-2}	0.008
KEGG Citrate Cycle TCA Cycle	10	1.64	9.22×10^{-3}	2.95×10^{-2}	10	1.39	2.94×10^{-2}	9.41×10^{-2}	0.02
MIPS 28S Ribosomal Subunit Mitochondrial	13	1.66	8.89×10^{-3}	3.02×10^{-2}	10	1.77	7.33×10^{-3}	2.35×10^{-2}	0.006

Table 5.5. Replicated gene set associations between European and African American data sets.
Gene sets listed in this table have an effective set size adjusted p-value at less than or equal to 0.10 in both European and African American data sets. Fisher's method ($-2 \times \ln[\text{p-value}_1 + \text{p-value}_2 \ldots \text{p-value}_x]$) was used to meta-analyse the p-values from both data sets.

The potential tissue specific expression of NEM gene sets was explored using NEM gene expression data from Pagliarini *et al.* (2008). Tissues of particular interest in this study were adipose, liver and skeletal muscle as they are major sites of insulin action. The number of NEM gene sets shared between these three tissues is presented in Table 5.6. A large proportion of all NEM gene are expressed in adipose tissue, which means for the purpose of identifying NEM gene sets, the full list and adipose-expressed list of NEM gene are for practical purposes the same. By contrast, using NEM gene expressed in skeletal muscle identified only 68% of the known NEM gene sets. The list of NEM genes expressed in the liver produced a unique NEM gene set not identified using the full NEM gene list, which was tested in the same way as the other NEM gene sets but it failed to reach significance.

	Complete MitoCarta	**Adipose**	**Liver**	**Muscle**
Complete MitoCarta	37	-	-	-
Adipose	37	37	-	-
Liver	31	31	32	-
Muscle	25	25	24	25

Table 5.6. Overlap between gene sets identified through the MSigDB using lists of expressed NEM genes.
Values in the table represent the total number of gene sets identified for each tissue (diagonal) and the number of NEM gene sets shared between the three tissues. The majority of gene sets identified from the NEM genes expressed in each of three tissues are common to all the three tissues.

With the exception of the NEMG988 and Wong mitochondrial gene sets due to their large sizes (Supplementary Table 5.1, 5.2, and 5.4), the results for other replicated NEM gene sets (Reactome pyruvate metabolism and citric acid TCA cycle, KEGG citrate cycle TCA cycle, and MIPS 28S ribosomal subunit mitochondrial) are presented in much greater detail in Tables 5.7-5.12. The tables provide the location estimate ($\hat{S}$) of the disease-susceptibility variant(s), its confidence interval, the size of the genomic region and the association statistic for the analytical window. The final two columns indicate whether the location estimate is within the gene coordinates and if the gene is located with the confidence interval for the analytical window.

A comparison between the European and African American results for any of the three replicated NEM gene sets reveal that gene set membership and size is not identical between the two samples. For example, the Reactome pyruvate and TCA cycle gene set has 17 genes (Supplementary Tables 5.1 and 5.2), all 17 are located within the genomic regions for the European sample (Supplementary Table 5.4) but only 12 are located within the genomic regions for the African American sample (Supplementary Table 5.4). The largest discrepancy in gene set membership and size between the European and African American samples is the Reactome pyruvate and TCA cycle gene set, in which there are four more genomic regions (Supplementary Table 5.4) in the European sample to which the genes of this gene set could be mapped to. This is a feature of mapping genes using a genetic rather than physical location and as explained previously, a genetic distance interpolated into a physical distance is usually shorter in the African American sample, resulting in a smaller genomic region to map genes into.

It can be seen from the χ^2 test statistics presented in Tables 5.7-5.12 that no NEM gene set contains unduly influential analytical windows. Taken as a whole, 68% of the analytical windows for the three replicated NEM gene sets have a significant ($\chi^2 > 3.84$) association statistic; individually, no NEM gene set has less than 50% of analytical window deemed significant. This demonstrates the mean χ^2 for each NEM gene set is result of multiple significant χ^2 values.

The general NEM gene sets as defined by MitoCarta and Wong *et al.* (2008) both contain genes located in genomic regions with very large χ^2 test statistics but these occurrences are infrequent. To illustrate this, the largest gene set, the MitoCarta list of approximately 1,000 NEM genes (378 genes in genomic regions, Supplementary Table 5.4), maps to 324 genomic regions in the European sample, with 317 of the 324 genomic regions below the 99th percentile for χ^2 test statistic distribution in this data set. The same gene set for the African American data sets maps to 248 genomic regions (277 genes in genomic regions, Supplementary Table 5.4), 241 of the 248 genomic regions fall below the 99th percentile for χ^2 test statistic distribution in this data set.

Reactome Pyruvate and TCA Cycle	Gene Coordinates	Chr	$\hat{S}$	Lower 95%CI	Upper 95%CI	Lower Genomic Region	Upper Genomic Region	χ^2	p-value	$\hat{S}$ Gene	CI Gene
SDHC	158.097-158.146	1q	158.292	158.129	158.405	158.088	158.375	0.27	$6.07x10^{-1}$	N	Y
SUCLG1	84.562-84.599	2p	84.880	84.816	84.919	84.301	85.052	13.68	$2.17x10^{-4}$	N	N
PDK1	173.246-173.315	2q	173.358	173.304	173.407	173.120	173.473	7.45	6.3510^{-3}	N	Y
SUCLG2	67.494-67.788	3p	67.831	67.800	67.865	67.576	67.867	7.83	$5.15x10^{-3}$	N	N
NNT	43.639-43.743	5p	43.273	43.273	43.301	42.897	43.650	21.78	$3.06x10^{-6}$	N	N
MDH2	75.322-75.341	7q	75.218	75.182	75.232	75.123	75.492	0.04	$8.52x10^{-1}$	N	N
DLD	107.125-107.166	7q	107.181	107.078	107.183	107.070	107.198	5.33	$2.09x10^{-2}$	N	Y
ADHFE1	67.505-67.546	8q	67.589	67.583	67.717	67.409	68.602	0.92	$3.38x10^{-1}$	N	N
PDHX	34.894-34.999	11p	34.815	34.814	34.815	34.631	34.992	6.90	$8.61x10^{-3}$	N	N
PDHX	34.894-34.999	11p	34.880	34.844	34.925	34.733	34.993	28.00	$1.22x10^{-7}$	N	Y
DLAT, *SDHD*	111.401-111.440; 111.463-111.496	11q	110.984	110.854	111.290	110.800	111.489	21.07	$4.43x10^{-6}$	N	N
LDHB	21.680-21.802	12p	21.710	21.706	21.868	21.636	21.909	6.34	$1.18x10^{-2}$	Y	Y
CS	54.952-54.980	12q	54.906	54.902	54.910	54.683	55.140	1.97	$1.61x10^{-1}$	N	N
L2HGDH	49.774-49.849	14	49.586	49.586	49.593	49.579	49.877	1.08	$2.98x10^{-1}$	N	N
IDH2	88.427-88.447	15	88.432	88.429	88.432	88.379	88.507	1.77	$1.84x10^{-1}$	Y	Y
BSG	0.522-0.534	19p	0.549	0.537	0.556	0.518	0.565	5.91	$1.51x10^{-2}$	N	N
ACO2	40.190-40.249	22	40.648	40.605	40.908	39.742	41.019	16.51	$4.83x10^{-5}$	N	N

Table 5.7. European results for the Reactome pyruvate metabolism and TCA cycle gene set.

Genomic coordinates (Mb) are human genome build 35/hg17. χ^2 and p-value represent the association of the analytical window with T2D. Abbreviations: Chr - Chromosome, CI - confidence interval, CI Gene - gene located with CI of analytical window, $\hat{S}$ - location estimate, $\hat{S}$ Gene - $\hat{S}$ within gene, Y - yes, N - no.

Reactome Pyruvate and TCA Cycle	Gene Coordinates	Chr	$\hat{S}$	Lower 95%CI	Upper 95%CI	Lower Genomic Region	Upper Genomic Region	χ^2	p-value	$\hat{S}$ Gene	CI Gene
SDHB	17.218-17.253	1p	17.183	17.177	17.231	17.078	17.278	0.47	$4.93x10^{-1}$	N	Y
FH	239.728-239.750	1q	239.775	239.771	239.778	239.745	239.828	18.82	$1.44x10^{-5}$	N	N
PDK1	173.128-173.198	2q	173.123	173.070	173.128	173.006	173.284	26.78	$2.28x10^{-7}$	N	Y
IDH1	208.809-208.839	2q	208.826	208.803	208.849	208.760	209.034	1.47	$2.25x10^{-1}$	Y	Y
SUCLG2	67.494-67.788	3p	67.768	67.767	67.817	67.697	67.845	51.82	$6.08x10^{-13}$	Y	Y
PDK4	95.051-95.064	7q	95.019	94.997	95.038	94.963	95.088	5.53	$1.87x10^{-2}$	N	N
DLD	107.319-107.359	7q	107.282	107.258	107.384	107.249	107.391	0.28	$5.97x10^{-1}$	N	Y
PDP1	94.939-95.007	8q	94.907	94.829	94.948	94.748	94.981	7.48	$6.24x10^{-3}$	N	Y
DLAT, *SDHD*	111.401-111.440; 111.463-111.496	11q	111.314	111.289	111.457	110.871	111.501	0.79	$3.74x10^{-1}$	N	Y
CS	54.952-54.980	12q	55.080	55.064	55.099	54.768	55.134	18.29	$1.90x10^{-5}$	N	N
L2HGDH	49.774-49.849	14q	49.740	49.607	49.797	49.593	49.912	9.84	$1.71x10^{-3}$	N	Y
SLC16A3	77.780-77.812	17q	77.758	77.631	77.783	77.627	77.783	1.99	$1.58x10^{-1}$	N	Y

Table 5.8. African American results for the Reactome pyruvate metabolism and TCA cycle gene set.
Genomic coordinates (Mb) are human genome build 36/hg18. χ^2 and p-value represent the association of the analytical window with T2D. Abbreviations: Chr - Chromosome, CI - confidence interval, CI Gene - gene located with CI of analytical window, $\hat{S}$ - location estimate, $\hat{S}$ Gene - $\hat{S}$ within gene, Y - yes, N - no.

KEGG TCA Cycle	Gene Coordinates	Chr	$\hat{S}$	Lower 95%CI	Upper 95%CI	Lower Genomic Region	Upper Genomic Region	χ^2	p-value	$\hat{S}$ Gene	CI Gene
SDHC	158.097-158.146	1q	158.292	158.129	158.405	158.088	158.375	0.27	6.07×10^{-1}	N	Y
MDH1	63.727-63.746	2p	62.798	62.732	62.822	62.636	63.855	27.42	1.64×10^{-7}	N	Y
SUCLG1	84.562-84.599	2p	84.880	84.816	84.919	84.301	85.052	13.68	2.17×10^{-4}	N	N
SUCLG2	67.494-67.788	3p	67.831	67.800	67.865	67.576	67.867	7.83	5.15×10^{-3}	N	N
MDH2	75.322-75.341	7q	75.218	75.182	75.232	75.123	75.492	0.04	8.52×10^{-1}	N	N
DLD	107.125-107.166	7q	107.181	107.078	107.183	107.070	107.198	5.33	2.09×10^{-2}	N	Y
DLAT, *SDHD*	111.401-111.440; 111.463-11.496	11q	110.984	110.854	111.290	110.800	111.489	21.07	4.43×10^{-6}	N	N
CS	54.952-54.980	12q	54.906	54.902	54.910	54.683	55.140	1.97	1.61×10^{-1}	N	N
IDH2	88.427-88.447	15	88.432	88.429	88.432	88.379	88.507	1.77	1.84×10^{-1}	Y	Y
ACO2	40.190-40.249	22	40.648	40.605	40.908	39.742	41.019	16.51	4.83×10^{-5}	N	N

Table 5.9. European results for the KEGG citrate cycle TCA cycle gene set.

Genomic coordinates (Mb) are human genome build 35/hg17. χ^2 and p-value represent the association of the analytical window with T2D. Abbreviations: Chr - Chromosome, CI - confidence interval, CI Gene - gene located with CI of analytical window, $\hat{S}$ - location estimate, $\hat{S}$ Gene - $\hat{S}$ within gene, Y - yes, N - no.

KEGG TCA Cycle	Gene Coordinates	Chr	$\hat{S}$	Lower 95%CI	Upper 95%CI	Lower Genomic Region	Upper Genomic Region	χ^2	p-value	$\hat{S}$ Gene	CI Gene
SDHB	17.218-17.253	1p	17.183	17.177	17.231	17.078	17.278	0.47	$4.93x10^{-1}$	N	Y
FH	239.728-239.750	1q	239.775	239.771	239.778	239.745	239.828	18.82	$1.44x10^{-5}$	N	N
MDH1	63.669-63.688	2p	63.712	63.494	63.759	63.062	63.922	48.81	$2.82x10^{-12}$	N	Y
IDH1	208.809-208.839	2q	208.826	208.803	208.849	208.760	209.034	1.47	$2.25x10^{-1}$	Y	Y
SUCLG2	67.494-67.788	3p	67.768	67.767	67.817	67.697	67.845	51.82	$6.08x10^{-13}$	Y	Y
DLD	107.319-107.359	7q	107.282	107.258	107.384	107.249	107.391	0.28	$5.97x10^{-1}$	N	Y
ACO1	32.375-32.441	9p	32.339	32.330	32.411	32.271	32.426	5.55	$1.85x10^{-2}$	N	Y
DLAT, SDHD	111.401-111.440; 111.463-111.496	11q	111.314	111.289	111.457	110.871	111.501	0.79	$3.74x10^{-1}$	N	Y
CS	54.952-54.980	12q	55.080	55.064	55.099	54.768	55.134	18.29	$1.90x10^{-5}$	N	N
PCK1	55.570-55.575	20q	55.596	55.595	55.598	55.574	55.604	0.1	$7.52x10^{-1}$	N	N

Table 5.10. African American results for the KEGG citrate cycle TCA cycle gene set.
Genomic coordinates (Mb) are human genome build 36/hg18. χ^2 and p-value represent the association of the analytical window with T2D. Abbreviations: Chr - Chromosome, CI - confidence interval, CI Gene - gene located with CI of analytical window, $\hat{S}$ - location estimate, $\hat{S}$ Gene - $\hat{S}$ within gene, Y - yes, N - no.

MIPS 28S Ribosomal Subunit	Gene Coordinates	Chr	$\hat{S}$	Lower 95%CI	Upper 95%CI	Lower Genomic Region	Upper Genomic Region	χ^2	p-value	$\hat{S}$ Gene	CI Gene
MRPS21	147.079-147.094	1q	146.809	146.626	146.814	146.528	147.330	3.04	$8.13x10^{-2}$	N	N
DAP3	152.471-152.522	1q	152.856	152.833	152.893	151.973	152.975	14.80	$1.19x10^{-4}$	N	N
MRPS23	140.207-140.559	3q	140.522	140.511	140.541	140.332	140.829	32.82	$1.01x10^{-8}$	Y	Y
MRPS27	71.551-71.652	5q	71.511	71.487	71.517	71.456	71.697	0.29	$5.88x10^{-1}$	N	N
MRPS33	140.159-140.168	7q	139.790	139.691	139.832	139.669	140.206	4.04	$4.44x10^{-2}$	N	N
MRPS28	80.994-81.105	8q	81.226	81.004	81.252	80.936	81.253	2.96	$8.54x10^{-2}$	N	Y
MRPS16	74.677-74.682	10q	74.125	73.724	74.380	73.402	75.367	45.00	$1.97x10^{-11}$	N	N
MRPS35	27.755-27.800	12p	27.810	27.744	27.857	27.695	27.917	1.76	$1.85x10^{-1}$	N	Y
MRPS31	40.201-40.243	13	40.038	40.031	40.159	39.968	40.590	22.40	$2.21x10^{-6}$	N	N
MRPS11	86.812-86.23	15	86.837	86.836	86.838	86.731	86.912	10.92	$9.50x10^{-4}$	N	N
MRPS34	1.762-1.763	16p	1.716	1.584	1.748	1.455	1.931	4.20	$4.05x10^{-2}$	N	N
MRPS23	53.272-53.282	17q	53.257	53.222	53.290	53.204	53.295	2.05	$1.53x10^{-1}$	N	Y

Table 5.11. European results for the MIPS 28S ribosomal subunit gene set.
Genomic coordinates (Mb) are human genome build 35/hg17. χ^2 and p-value represent the association of the analytical window with T2D. Abbreviations: Chr - Chromosome, CI - confidence interval, CI Gene - gene located with CI of analytical window, $\hat{S}$ - location estimate, $\hat{S}$ Gene - $\hat{S}$ within gene, Y - yes, N - no.

MIPS 28S Ribosomal Subunit	Gene Coordinates	Chr	$\hat{S}$	Lower 95%CI	Upper 95%CI	Lower Genomic Region	Upper Genomic Region	χ^2	p-value	$\hat{S}$ Gene	CI Gene
MRPS15	36.694-36.703	1p	36.630	36.630	36.634	36.580	36.705	17.01	$3.72x10^{-5}$	N	N
MRPS21	148.533-148.548	1q	148.478	148.432	148.484	148.405	148.755	0.84	$3.59x10^{-1}$	N	N
MRPS23	140.207-140.559	3q	140.477	140.468	140.482	140.344	140.525	6.6	$1.02x10^{-2}$	Y	Y
MRPS33	140.352-140.361	7q	140.171	140.000	140.293	139.927	140.454	32.41	$1.25x10^{-8}$	N	N
MRPS2	137.532-137.536	9q	137.523	137.510	137.529	137.510	137.564	0.06	$8.06x10^{-1}$	N	N
MRPS2	137.532-137.536	9q	137.564	137.529	137.587	137.523	137.569	0.46	$4.98x10^{-1}$	N	Y
MRPS16	74.677-74.682	10q	75.031	74.399	75.074	74.323	75.077	44.48	$2.57x10^{-11}$	N	Y
MRPS11	86.812-86.823	15q	86.837	86.836	86.839	86.747	86.942	5.55	$1.85x10^{-2}$	N	N
MRPS34	1.762-1.763	16p	1.782	1.761	1.865	1.541	1.936	42.42	$7.36x10^{-11}$	N	Y
MRPS7	70.769-70.774	17q	70.656	70.622	70.815	70.606	70.928	6.540	$1.05x10^{-2}$	N	Y

Table 5.12. African American results for the MIPS 28S ribosomal subunit gene set.

Genomic coordinates (Mb) are human genome build 36/hg18. χ^2 and p-value represent the association of the analytical window with T2D. Abbreviations: Chr - Chromosome, CI - confidence interval, CI Gene - gene located with CI of analytical window, $\hat{S}$ - location estimate, $\hat{S}$ Gene - $\hat{S}$ within gene, Y - yes, N - no.

Discussion

This Chapter has presented evidence for the identification of five mitochondrial gene sets associated with T2D in European and African American populations. The approach employed in this study is unique for two reasons; the first is the use of multi-marker test of association as opposed to the single-marker test. Secondly, due to the nature of data that is generated from this type of genetic analysis based upon analytical windows across the genome, it is possible to map entire gene sets onto the genome to perform the gene set analysis. Publically available gene set analysis tools such as GSEA (Subramanian *et al.*, 2005), MAGENTA (Segre *et al.*, 2010), DAVID (Huang da *et al.*, 2009) are designed with gene expression microarray or single SNP data in mind and therefore are unsuitable for this mapping procedure.

Of the five NEM gene sets potentially identified, two are lists of NEM genes (Wong mitochondria gene module and NEMG988), whilst the remaining three all relate to specific biological processes involved in ATP production - pyruvate metabolism, the TCA cycle, and mitochondrial protein translation. These connected biological pathways encompass all cellular ATP production, which is a key determinant of glucose-stimulated insulin secretion (Chapter 1: Insulin and Glucose Action). The pyruvate metabolism and the TCA cycle pathways have been identified in a metabolic approach to characterise the amplifying pathway of glucose-stimulated insulin secretion in glucose-responsive and unresponsive pancreatic beta cells (Spegel *et al.*, 2011). The authors were able to demonstrate glucose unresponsiveness was due to the decoupling of glycolysis and the TCA cycle, which could be circumvented by stimulating the TCA directly, resulting in increased oxygen consumption and insulin secretion of the glucose unresponsive cell line.

In extreme diabetic conditions (>25 mM glucose), as is seen in diabetic ketoacidosis and the hyperglycaemic hyperosmolar state, the pancreatic beta cells might be expected to maximise secretion of insulin in an attempt to restore normoglycaemia. However, the opposite is true – insulin secretion is reduced and this reduction is in part due to suppression of the TCA cycle (Wallace *et al.*, 2013).

As for the mitochondrial translation gene set, it could be proposed that alterations in the mitochondrial ribosomal subunit affect the translation of mitochondrial mRNA into the polypeptide sequence, which leads to compromised or inefficient mitochondrial protein synthesis. Mutations in the transfer RNA units of mitochondrial can lead to certain type of diabetes as discussed in Chapter 1. Thus, some (Jacobs and Turnbull, 2005, Scheper *et al.*, 2007, Smits *et al.*, 2010) have suggested the mitochondrial ribosomal is not impervious to similar deleterious mutations, and draw attention to an example documented in Drosophila (Toivonen *et al.*, 2001). In this example, the mutation in the *MRPS12* gene that encodes one of proteins that form the small mitochondrial subunit disrupts the stability of the subunit and results in complete loss of oxidative phosphorylation activity of the mitochondria.

The 13 protein-encoding genes in the mitochondrial genome encode critical components of the oxidative phosphorylation complexes. What is the significance of this allocation of genes within the mitochondrial genome? One theory (Allen, 2003), the co-location for redox regulation (CORR), suggests mitochondria must have the ability to swiftly regulate the production of damaging reactive oxygen species (ROS) as a consequence of oxidative phosphorylation. It is thought that the electron carriers of this pathway sense the redox state of the mitochondrion and can down regulate turnover through mitochondrial protein synthesis restriction (Escobar Galvis *et al.*, 1998).

If the CORR hypothesis is correct, the reason why these genes have not transferred to the nucleus is because of a need for an immediate fail-safe switch at the site of ROS production to limit the damage inflicted upon the cell by these molecules. Alternatively, the presence of a small number of oxidative phosphorylation genes in the mitochondrial genome may simply reflect the fact that the transfer of organelle genes to the nucleus takes a very long time (Keeling and Palmer, 2008, Leister, 2005). The latter explanation is negated by the fact that nearly all organisms that rely on aerobic respiration retain a mitochondrial genome, which indicates a shared evolutionary constraint and not merely that there has not yet been enough insufficient time to transfer the mitochondrial genes to the nucleus.

Previous studies (Mootha *et al.*, 2003, Segre *et al.*, 2010, Perry *et al.*, 2009) have used gene set methods in relation to T2D. Differential expression of nuclear genes that encode members of the oxidative phosphorylation pathway were shown by Mootha *et al.* (2003) to be associated with T2D; we were unable to replicate this finding using our data in that the Mootha OXPHOS gene set (Supplementary Table 5.1) did not reach nominal significance in either European or African American populations. It is likely our analysis for this particular gene set was underpowered since approximately two-thirds of the genes from this gene set were discarded, as they were not located within the defined genomic regions.

Segre *et al.* (2010) could not find evidence of association for genetic variants in the oxidative phosphorylation gene set for T2D and related phenotypes; their explanation for the result is that down regulation of the oxidative phosphorylation gene set is a consequence of T2D rather than contributory. This is certainly reasonable, as studies on the expression of oxidative phosphorylation genes have been cross-sectional in design, so causality cannot be unambiguously inferred. In addition, expression of this gene set in T2D appears to be coordinated across different tissues; it is down regulated in skeletal muscle (Mootha *et al.*, 2003, Skov *et al.*, 2007) and pancreatic beta cells (Olsson *et al.*, 2011b), whereas it is up regulated in the liver (Buchner *et al.*, 2011).

With this in mind, a more likely explanation for why Segre *et al.* (2010) did not observe an association between oxidative phosphorylation genes and T2D, where as the gene expression studies persuasively show a relationship does exist, is the distinction between SNP data and gene expression data. Assuming the effect of somatic mutation is negligible, the genotype at a locus does not change over an individual's lifetime nor does it vary between cells of different tissues. On the other hand, the expression of a gene at the level of mRNA transcript abundance can vary by orders of magnitude between cells of different tissues and over time. Tissue specific changes in gene expression are not going to be observed with SNP data.

Perry *et al.* (2009) offers slightly more commensurability with the study presented here in that both use data from the WTCCC1 T2D study (Wellcome Trust Case Control, 2007). Of the 26 gene sets that were nominally significant at an alpha of 5%

presented in Perry *et al.* (2009), only one (KEGG pyruvate metabolism) also appears in this study. A significant association for this gene set is observed in the European data set after adjusting for effective gene set size but was not replicated in our African American sample.

Elbers *et al.* (2009) have expressed concerns regarding gene set analysis and its interpretation. Pertinent to this study are the influence of large genes, large gene sets, and representativeness of gene sets from publically available databases. They suggest permutation of gene set membership would lessen the bias caused by large genes within gene sets, which is one of the reasons why we incorporated permutation into our analyses.

The tendency of large gene sets to produce large test statistics under the null has been observed in this study, which is the same relationship observed by both Elbers *et al.* (2009) and Perry *et al.* (2009). As explained above, this has been addressed by some means such as down-weighting the p-value for a gene set by its effective gene set size. Thirdly, interpretation of the results of a gene set analysis needs to be approached with caution as the interpretation is only as good as the quality of the data being used - in particular gene set definition. For this reason, inclusion of MSigDB gene sets for analysis was restricted to those that had been manually curated. While these gene sets are of high quality based upon well-documented evidence, the limitation of curated gene sets is that, by definition, these are restricted to well-characterised gene pathways.

Gene sets are seldom defined in an identical manner between different sources and where they are, these are likely to be small gene sets that represent extensively studied biology pathways. It is far more likely that a gene-set from two independent databases or defined under different experimental conditions within the same database will share some proportion of genes, but will also will contain genes unique to each source/condition. The hope is enough genes are common to both in order to give consistent results when analysed.

Many of the 38 gene sets that were analysed in this study are ostensibly based on shared biological pathways; for example, four of the 38 gene sets (Supplementary

Table 5.1) attempt to single out the tricarboxylic acid (TCA) cycle, a core biological pathway for aerobic respiration of glucose. These four definitions of the TCA cycle, each from a different source (Biocarta, KEGG, Reactome, and Mootha *et al.* (2003)), vary in the number of genes from 8 to 32. If we also include gene sets that represent a combination of the TCA cycle and another biological pathway, then there are six TCA cycle gene sets, varying in size from 8 to 141 genes. Figure 5.5 illustrate that despite the inclusion of multiple gene sets for the same biological pathway, the gene sets are independent entities, in other words, the proportion of genes shared between any two gene sets is small.

The nominally replicated gene sets are considered to be of most interest are because they represent specific biological pathways (pyruvate metabolism and the TCA cycle) or function (subcomponent of mitochondrial translation). In total, these three mapped gene sets contain 31 unique genes (Supplementary Tables 5.5 – 5.8). None of the mapped genes in these mitochondrial gene sets have been implicated in T2D or related phenotypes through GWA studies according to the NHGRI catalogue (Hindorff *et al.*, 2009), accessed 08/04/13.

A search of the Expression Atlas and the Mouse Genome Informatics (MGI) resources was undertaken to find any evidence that would connect these genes with T2D. Twelve of the 31 genes show expression evidence (Supplementary Table 5.5 – 5.8) that potentially relates them to T2D; insulin secretion (*PDK1*, *IDH2*, *NNT*), impaired glucose tolerance (*NNT*), coupling electron transport and oxidative phosphorylation (*MRPS22*), adipogenesis (*MDH1*, *ADHFE1*), hyperglycaemia-induced inflammatory response (*BSG*), hepatic response to high fat diet (*ACO2*, *MRPS22*), lipid-storage (*MRPS35*) and gestational diabetes (*MRPS21*, *MRPS27*).

The results of Table 5.5 are strictly limited to gene sets that have an adjusted p-value of less than 0.10 in both European and African American data sets. If each population is considered separately, then using the same alpha threshold there are 13 observed gene sets for the European sample and coincidentally, 13 gene sets for the African American sample, five of which appear in both data sets as listed in Table 5.5.

As a subsidiary research-method question relating to gene set analysis, the data was also analysed using an overrepresentation method (Fisher's exact test) in order to be compared with results obtained for the set-based (two sample t-test) method. The standard two-sample t-test produced a greater number of potentially replicated gene sets for two unrelated case/control samples, compared to the Fisher's exact test. However, the only gene sets from the overrepresentation analysis that replicated between the two samples (unadjusted p-value ≤ 0.10) were the largest (effective gene set size) of the 38 gene sets analysed. The t-test provided a much more varied set of gene sets, some as small as just 10 genes returned unadjusted p-values below 0.01.

Study Strengths and Limitations

As a direct consequence of the gene mapping strategy some gene sets were poorly represented at the stage of analysis. The average reduction in gene set size was approximately 70%. As explained in the Materials and Methods (Chapter 5), each genomic region utilised the location estimate for each analytical window, which can be interpreted as the most likely location for a causal T2D susceptibility variant within an analytical window. The location estimate for all analytical windows was used, irrespective of the association of the window with T2D and extended it into a region by adding a constant genetic distance to both sides of the estimate. Genetic distance is a function of recombination, thus when interpolated to physical distance, the genomic regions will vary in size according to local recombination.

In the European sample, 98.56% (4,242 / 4,304) of the genomic regions are smaller than the analytical window they originate from. As a genome-wide average, the location estimate regions are 58.70% smaller than their parent analytical window in terms of the physical length. The African American data show a similar trend whereby 99.56% (4,517 / 4,537) of the genomic regions are smaller than their parent analytical windows and the genomic regions are on average 70.26% smaller than the analytical windows.

By centring on the location estimate and shortening the region of interest, specific statements can be made about the role of nuclear-encoded mitochondrial genes in T2D susceptibility. As the location estimate regions are smaller than the analytical windows, whenever a nuclear-encoded mitochondrial gene set collectively maps to within these regions, then by design, they are within close proximity of the T2D association signal if one is present. Had the analytical windows been used for gene set mapping, it would be expected that many more genes of mitochondrial gene sets could have been successfully mapped, but would ignore information about location of the susceptibility variant. For example, a mitochondrial gene and a T2D susceptibility variant may co-localise to an analytical window, but possibly at opposite ends of the window.

A potential limitation may be that our 100,000 permutations of gene set membership may not have been sufficient to exhaust all possible combinations of gene set membership for each gene set. We chose this figure as a compromise between the possibility of a particular gene set appearing with identical members in one of the permutations and the computational time required to run the analysis, which is fairly substantial for a total of 3.8 million permutations (100,000 permutations for 38 gene sets).

Concluding Chapter

In the concluding chapter I provide a brief outline of my main research findings and discuss these in relation to the literature, the strengths and weaknesses of my work, and finally, building upon this research, I discuss what I think might be worthwhile future research questions to pursue.

This thesis begins by evaluating various measures of adiposity and their effect on cardiovascular and metabolic co-morbidities associated with T2D; as T2D is the main disorder studied in this thesis, it is important to understand how adiposity and body fat distribution relate to disease. Almost ubiquitously, BMI is used as the measurement for defining obesity. Its inappropriateness is discussed in the introduction of Chapter 3. Despite its wide usage, the measurement tells the investigator nothing about the amount or location of fat an individual carries - the correlation between BMI and percentage body fat or between BMI and lean (muscle) mass are similar (Romero-Corral *et al.*, 2008).

Rather than use generalised measures of adiposity as a height-corrected measurement of body mass, a better solution in assessing the risk conferred is to actually measure the amount and location of fat directly. Fortunately, the TwinsUK resource has multiple adiposity phenotypes on the same individual, including computer-assisted and manually measured, for a large number of individuals. This feature allows useful comparisons to be made across adiposity phenotypes.

For T2D, abdominal fat and more specifically, intra-abdominal (visceral) fat, has received considerable interest as a contributory risk factor for the disease. Partly due to the expense, very few individuals of TwinsUK have a direct, computed tomography measurement of their visceral fat. Nevertheless, the literature contains a wealth of information that consistently indicate that various proxy measures of visceral fat can be created by the judicious combination of relevant measurements. This study was one of the first to systematically assess the validity and reliability of DXA and anthropometry-based measures of visceral fat. This proxy measure for a large sample of twins was also used to assess the role of visceral fat in contributing to morbidity relative to other measures of body fat distribution.

For this study it was possible to assess a number of previously reported visceral fat estimation models using TwinsUK data as shown in the results of Chapter 3. A heritability study was conducted that showed 57% of the phenotypic variance could be attributed to additive genetic variance, which is consistent with other heritability studies using the classical twin model, and to my knowledge, this study is the largest heritability study of visceral fat conducted. The bivariate analysis between the estimate of visceral fat and DXA total abdominal fat, showing a genetic component specific to estimated visceral fat and not shared by DXA total abdominal fat, also provided additional evidence of measurement validation.

Having a measure of visceral fat made it possible to better address the issue of body fat distribution in relation to morbidity. This is important because a number of studies demonstrate that not all fat depots behave in the same way – fat appears to be safely stored in subcutaneous regions, whilst being particularly detrimental when stored close to the internal organs. The message from these studies is that location of fat storage is more important that the actual quantity of fat. One of the main findings of Chapter 3 was, almost without exception, visceral fat was the biggest independent risk factor for the cardiovascular and metabolic morbidities that were investigated. This suggests weight loss programs might perhaps be more effective in reducing the burden of these conditions on the health service, if they tracked intra-abdominal fat rather than waist circumference or BMI.

The focus of Chapter 4 was a candidate locus study for T2D. The starting point of the study was to attempt to replicate a previously reported association between a SNP in the *PARL* gene and fasting insulin level. Approximately 3,000 individuals from TwinsUK were genotyped for this SNP along with 26 additional SNPs to improve coverage in this genomic region. Despite this effort, the SNP-insulin association could not be replicated nor were any SNPs in the *PARL* gene associated with fasting insulin or glucose based upon a single SNP test of association.

Analysis of the same genomic region using a fine-scale genetic map and a multi-marker test of association identified a strong association signal with T2D in the neighbouring gene, *ABCC5*. The strongest association was seen for the WTCCC1 sample, with both the African American and TwinsUK (Human610-Quad sample)

replicating the statistical association with an identical location estimate. In addition to SNP data, a subset of TwinsUK also had gene expression data, which meant an eQTL analysis could be performed to assess whether the T2D/IGR association, was perhaps in part, driven by the expression of *ABCC5*. A very strong signal was found for *ABCC5* gene expression in subcutaneous adipose tissue, both confirming a role for *ABCC5* and the same genomic location estimate for the functional variant (at 185,136Kb). This suggests the susceptibility variant is itself an eQTL for *ABCC5* and confers risk of disease via increased expression levels of *ABCC5*.

Relative to gene expression in LCL or skin, subcutaneous adipose tissue was the primary tissue of interest for T2D in relation to the expression patterns of *PARL* and *ABCC5*. In subcutaneous adipose tissue, only one eQTL was identified and was strongly associated with the expression of *ABCC5*, whereas the other tissues provided a less clear picture with several eQTLs that act both locally and in the neighbouring gene. *ABCC5* (but not *PARL*) expression in subcutaneous adipose tissue was positively associated with T2D and related traits, including the estimated measure of visceral fat from Chapter 3, for which a one standard deviation increase in *ABCC5* expression value is associated with a 30 cm^2 increase in visceral fat area.

Collectively the results presented in Chapter 4 provide firm evidence for *ABCC5* to be involved in the aetiology of T2D. The gene has received little attention from large T2D consortia, which primarily advocate ever-larger sample sizes, but still use a single SNP test of association incurring a large multiple testing penalty.

In the WTCCC1 T2D study (Wellcome Trust Case Control, 2007), the p-value for the most significantly associated SNP in the *PARL*/*ABCC5* region was only 10^{-3}. By contrast, for the same region and the same data, the multi-marker test provided a model p-value of 10^{-6} and a functional variant location estimate in intron 26 of *ABCC5* (185,360Kb). At a conventional genome-wide significance threshold (500K SNPs, p-value < 10^{-7}), neither result would be considered to be a significant association. However, the multi-locus Malécot model (with a model fit p-value of 10^{-6}) provides significant evidence of association for a candidate gene study, using an alpha threshold of 10^{-4}. Indeed, using a Malécot model approach, a genome-wide alpha threshold of 10^{-5} is also appropriate, after accounting for multiple testing

involved (Elding *et al.*, 2013). The genome-wide significance threshold for the multi-marker approach (10^{-5}) is larger than the conventional GWAS threshold of 10^{-7}, because the multi-marker analysis divides the entire genome into nearly 5,000 analytical windows of equal genetic distance and tests each window - not each SNP. This is equivalent to approximately 5000 independent tests of association (Bonferroni corrected p-value = 0.05/ 5000 = 10^{-5}). The impact of this is a genome-wide significance threshold of 10^{-5} – two orders of magnitude larger than genome-wide single SNP tests. Taken together, the combined effect of testing SNPs in the aggregate and reducing the number of tests being performed explains why an association between T2D and *ABCC5* is observed in this study, which uses the data from the WTCCC1 T2D study, but not in the original WTCCC1 T2D study.

Aside from statistical evidence, there is also intriguing *ABCC5* functional evidence documented in the literature – Nowicki *et al.* (2008) showed that destruction of pancreatic beta cells in rats combined with a high fat diet reduces protein abundance of Abcc5 by almost 100% in hepatocytes. It is interesting to speculate how removing insulin on a high fat diet causes this effect. Other than the study cited above, to my knowledge there is no literature to explain the relationship between insulin and *ABCC5* and how a high fat diet modulates their interaction.

Chapter 4 began by looking the *PARL* gene, which is mitochondrial gene encoded by nuclear DNA. As explained in the introduction, mitochondria are an essential component in glucose-stimulated insulin secretion. Hence, the fact that very few susceptibility variants have been identified near or within mitochondrial or nuclear-encoded mitochondrial genes is somewhat counter-intuitive, and probably reflects inadequate study designs and/or inappropriate statistical methods.

The analysis of biological pathways rather than variants in individual genes is a relatively new approach for identifying additional disease susceptibility genes and pathways. Pathway analysis first grew in popularity in its application to genome-wide expression studies and then was adapted for SNP data analysis. One of the reasons for its appeal is that sets of genes are based on prior biological knowledge rather than identifying a single gene as part of a hypothesis-free approach. Moreover, it captures the concept that for complex traits, it is expected that susceptibility loci do not act in

isolation. The results of the mitochondrial gene set analysis in Chapter 5 show NEM pathways appear enriched for signals of T2D association - pyruvate metabolism, the TCA cycle, and mitochondrial translation. These pathways all relate to the generation of cellular energy, the key determinant of insulin secretion by the pancreatic beta cells and therefore of major interest in the treatment of T2D. Currently the only oral anti-hyperglycaemic that targets the pancreatic beta cells are sulfonylureas, which act by closing the ATP-sensitive potassium channel. Increasing the availability of ATP in these cells by enhancing ATP generation potentially provides an additional method to control hyperglycaemia at the pancreatic beta cell.

Future work

As discussed in the end of Chapter 3, there is considerable utility for a non-CT method of measuring visceral fat in large research data resources such as the UK Biobank (Palmer, 2007). In effect, such a resource would be able to acquire both DXA-assessed abdominal already being collected and at no additional cost, a DXA-based visceral fat estimate using routine and inexpensive measures of abdominal anthropometry, such as waist circumference or sagittal depth, and age.

Since a large proportion of the TwinsUK individuals that have visceral fat data also have genome-wide SNP data available (Chapter 2: TwinsUK Sample), it would be possible perform a standard GWA scan for visceral fat in this sample. However, considering that the estimate is an inexact proxy with approximately 20% measurement error compared to CT, in the context of the multiple testing burden of GWA scans, it would be unwise to perform the GWA scan for two main reasons. Any significant hits would have to be validated in an independent study (which is true for all GWA studies) but more substantially, the fact it is an imperfect proxy measurement, will likely result in an inflated Type I error rate compared to most GWA studies for common disease.

To date, only one GWA study has been performed for visceral fat (Fox *et al.*, 2007) with a large with a sample size of approximately 10,000 participants drawn from three population based studies - the Framingham Heart Study (FHS), a subset of this - the Family Heart Study, and the Health, Aging, and Body Composition (Health ABC) Study. Only one SNP in the *THNSL2* gene was identified by Fox *et al.* (2007) and the association was only seen in females, which led the authors to the somewhat anodyne conclusion that many more susceptibility loci remain to be identified. The FHS and the Health ABC study appear to allow external requests for genotype and phenotypic data or at the very least, GWA scan p-values; it would be interesting to re-analyse their data to perform a genome-wide multi-marker association test in the hope of identifying more susceptibility variants for visceral fat.

Functional work is required to understand how *ABCC5* is mechanistically implicated in the aetiology of T2D and how this gene might provide broader insights into the disease process. To date, the *ABCC5* protein has largely received attention in its role of conferring resistance to thiopurine anticancer drugs. The analysis presented in this thesis is certainly consistent with a contributory role of *ABCC5* in the development of T2D. The fact that *ABCC5* encodes a transmembrane protein also opens up the possibility of it being a drug target in T2D treatment. The work presented in this thesis suggests a potentially direct causal role for *ABCC5* in the development of T2D, specifically – the association of fasting insulin levels, visceral fat and T2D with *ABCC5* expression levels. In addition, the destruction of the insulin secreting cells of the pancreas combined with a high diet is able to eliminate hepatic *ABCC5* expression (Nowicki *et al.*, 2008) also illustrates that in more extreme circumstances, the loss of *ABCC5* function can also be a consequence of T2D.

The gene set analysis (Chapter 5) provides potentially novel insights into the aetiology of T2D, implicating biological pathways that relate to the production of ATP through cellular respiration (glycolysis, the TCA cycle, and indirectly, oxidative phosphorylation through mitochondrial protein synthesis [mitochondrial ribosome translational unit, 28S]). For glycolysis and the TCA cycle, a fruitful way to confirm these results is to adopt a similar strategy to that used by Spegel *et al.* (2011). Their study showed that glucose unresponsiveness of pancreatic beta cells could be explained by the decoupling of products from pyruvate metabolism to the TCA cycle. Measuring the individual metabolites of these pathways in relevant tissues might reveal differences between normoglycaemic and T2D individuals and potentially therapeutic targets.

For the mitochondrial translation gene set, there is a direct connection with glucose-stimulated insulin secretion. As discussed in Chapter 1, it was noted that the 13 protein coding genes on the mitochondrial genome are essential components of the electron transport system, which performs oxidative phosphorylation and are ultimately responsible for the generation of ATP. The increase in ATP/ADP ratio is an early and key event in glucose-stimulated insulin secretion from the pancreatic beta cells. A systematic deletion or knock down of each nuclear–encoded mitochondrial ribosomal gene could be performed and ATP/ADP ratio be measured directly (Berg *et*

al., 2009, Imamura *et al.*, 2009, Bugger *et al.*, 2009, Chinopoulos *et al.*, 2009). The results from this type of study could provide important insights into beta cell function in relation to T2D and also, in an evolutionary context, the relative importance and role of genes that have been retained in the mitochondrion compared to those that have translocated from the mitochondrion to the nucleus.

Supplementary Tables

	VAT	DXA	BMI	WC
DXA	0.90			
BMI	0.78	0.77		
WC	0.96	0.83	0.79	
Age	0.42	0.26	0.17	0.19

Supplementary Table 3.1. Correlation and colinearity between measures of adiposity for study sample (n = 3,457).

Pearson product-moment correlation coefficient is presented. Abbreviations: BMI - body mass index, DXA - total abdominal fat measured using dual-energy X-ray absorptiometry, WC - tape-measured waist circumference (cm) taken at DXA scan visit.

T2D		OR	SE	z	p-value	95% CI		Model-fit		
						Lower	Upper	Pseudo R^2	χ^2 (df)	p-value
Full model								0.08	-	-
	VAT	2.50	0.62	3.7	$2.2x10^{-4}$	1.54	4.05			
	DXA	1.00	0.17	0.0	$9.9x10^{-1}$	0.71	1.41			
	BMI	0.84	0.13	-1.2	$2.3x10^{-1}$	0.62	1.13			
	Age	1.02	0.01	1.5	$1.3x10^{-1}$	0.99	1.04			
Drop VAT								0.07	14.4 (1)	$1.5x10^{-4}$
	DXA	1.53	0.17	3.8	$1.5x10^{-4}$	1.23	1.92			
	BMI	1.22	0.14	1.8	$7x10^{-2}$	0.98	1.52			
	Age	1.04	0.01	3.4	$6.7x10^{-4}$	1.02	1.06			
Drop DXA								0.08	0.00 (1)	$9.9x10^{-1}$
	VAT	2.51	0.43	5.3	$1.2x10^{-7}$	1.79	3.52			
	BMI	0.84	0.13	-1.2	$2.3x10^{-1}$	0.62	1.13			
	Age	1.02	0.01	1.5	$1.3x10^{-1}$	1.00	1.04			
Drop BMI								0.08	1.9 (1)	$1.7x10^{-1}$
	VAT	2.08	0.40	3.8	$1.5x10^{-4}$	1.43	3.02			
	DXA	1.00	0.18	0.01	$9.9x10^{-1}$	0.71	1.41			
	Age	1.02	0.01	1.91	$6x10^{-2}$	1.00	1.05			
Drop DXA & BMI								0.08	1.9 (2)	$3.9x10^{-1}$
	VAT	2.08	0.18	8.5	$1.9x10^{-17}$	1.76	2.47			
	Age	1.02	0.01	2.0	$5x10^{-2}$	1.00	1.05			
Drop VAT & BMI								0.07	18.3 (2)	$1.1x10^{-4}$
	DXA	1.81	0.14	7.9	$2.8x10^{-15}$	1.56	2.10			
	Age	1.04	0.01	3.2	$1.4x10^{-3}$	1.01	1.06			
Drop VAT & DXA								0.06	27.0 (2)	$1.4x10^{-6}$
	BMI	1.65	0.12	6.9	$5.2x10^{-12}$	1.43	1.90			
	Age	1.04	0.01	3.8	$1.5x10^{-4}$	1.02	1.06			

Supplementary Table 3.2 Type 2 diabetes (T2D) multiple regression analyses.
The likelihood ratio test (LRT) statistic p-value is presented to assess the decline in model fit when one or more measures of adiposity are removed from the full model. Note that dropping total abdominal fat (DXA) and/or body mass index (BMI) does not diminish the model fit, while dropping visceral adipose tissue (VAT) area does (p-value < 0.05).

Type 2 diabetes		β	SE	t	p-value	95% CI		Model-fit	
						Lower	Upper	Pseudo R^2	p > F
A: *Visceral fat residuals*								$7x10^{-4}$	0.38
T2D' ~									
	DXA'	0.00	0.04	0.07	0.95	-0.08	0.09		
	BMI'	-0.05	0.04	-1.39	0.17	-0.12	0.02		
B: *DXA & BMI residuals*								$5x10^{-3}$	$1x10^{-4}$
T2D"~									
	VAT"	0.23	0.06	3.89	$2.8x10^{-4}$	0.11	0.34		
Where residuals are for:									
A	T2D' = T2D ~ VAT + Age								
	DXA' = DXA ~ VAT+ Age								
	BMI' = BMI ~ VAT+ Age								
B	T2D" = T2D ~ DXA + BMI + Age								
	VAT" = VAT ~ DXA + BMI + Age								

Supplementary Table 3.3. Residuals analysis for type 2 diabetes.

No residual association remains (p-value = 0.38) between T2D and total abdominal fat (DXA) and body mass index (BMI), when the residuals for T2D, DXA and BMI are taken in relation to visceral adipose tissue (VAT) area (**A**). The residuals for T2D are on the logit scale, while the DXA and BMI residuals use ordinary least squares regression. By contrast, a small but statistically significant residual association remains between T2D and VAT (**B**), when the residuals for T2D and VAT are taken in relation to DXA, BMI and age. Despite co-linearity between VAT, DXA and BMI (see Supplementary Table 3.1), these results strongly indicate that the association between T2D and 3 measures of adiposity is primarily with VAT area.

Probe name	Probe location (B36)	Gene	Ensembl Transcipt IDs
Ilmn_2341467	185030030 – 185030079	*PARL*	*PARL*-001, *PARL*-002, *PARL*-004, *PARL*-007, *PARL*-008, *PARL*-009, *PARL*-010
Ilmn_1731354	185030034 – 185030083	*PARL*	*PARL*-001, *PARL*-002, *PARL*-004, *PARL*-007, *PARL*-008, *PARL*-009, *PARL*-010
Ilmn_2257665	185042790 – 185042839	*PARL*	*PARL*-001, *PARL*-004, *PARL*-006, *PARL*-008, *PARL*-009, *PARL*-010
Ilmn_1706531	185120566 – 185120615	*ABCC5*	*ABCC5*-001, *ABCC5*-007
Ilmn_1651964	185184355 – 185184404	*ABCC5*	*ABCC5*-003, *ABCC5*-004, *ABCC5*-005, *ABCC5*-006
Ilmn_2302358	185188345 – 185188394	*ABCC5*	*ABCC5*-001, *ABCC5*-003, *ABCC5*-004, *ABCC5*-005, *ABCC5*-006, *ABCC5*-007, *ABCC5*-009, *ABCC5*-010, *ABCC5*-201

Supplementary Table 4.1. Microarray gene expression probes tag a population of mRNA transcripts for *PARL* and *ABCC5*.

Index	Gene Set
1	Biocarta ETC pathway
2	Biocarta Kreb pathway
3	Galluzzi Prevent Mitochondial Permeabilization
4	KEGG Beta Alanine Metabolism
5	KEGG Butanoate Metabolism
6	KEGG Citrate Cycle TCA Cycle
7	KEGG Fatty Acid Metabolism
8	KEGG Glyoxylate and Dicarboxylate Metabolism
9	KEGG Limonene and Pinene Degradation
10	KEGG One Carbon Pool by Folate
11	KEGG Oxidative Phosphorylation
12	KEGG Parkinsons Disease
13	KEGG Propanoate Metabolism
14	KEGG Pyruvate Metabolism
15	KEGG Valine Leucine and Isoleucine Degradation
16	MIPS 28S Ribosomal Subunit Mitochondrial
17	MIPS 39S Ribosomal Subunit Mitochondrial
18	MIPS 55S Ribosome Mitochondrial
19	MIPS F1F0 ATP Synthase Mitochondrial
20	Mootha FFA Oxydation
21	Mootha Human MitoDB6 2002
22	Mootha Mitochondria
23	Mootha TCA
24	Mootha VOXPHOS
25	Reactome Branched Chain Amino Acid Catabolism
26	Reactome Citric Acid Cycle TCA Cycle
27	Reactome Formation Of ATP by Chemiosmotic Coupling
28	Reactome Mitochondrial Fatty Acid Beta Oxidation
29	Reactome Mitochondrial Protein Import
30	Reactome Mitochondrial tRNA Aminoacylation
31	Reactome Pyruvate Metabolism
32	Reactome Pyruvate Metabolism and Citric Acid TCA Cycle
33	Reactome Regulation of Pyruvate Dehydrogenase PDH Complex
34	Reactome Respiratory Electron Transport
35	Reactome Respiratory Electron Transport ATP Synthesis by Chemiosmotic Coupling and Heat Production by Uncoupling Proteins
36	Reactome TCA Cycle and Respiratory Electron Transport
37	Wong Mitochondria Gene Module
38	NEMG988

Supplementary Table 5.1 List of nuclear-encoded mitochondrial gene sets analysed.

Index	Number of Genes	Number of Mapped Genes
1	12	9
2	8	8
3	22	22
4	22	21
5	34	33
6	32	28
7	42	41
8	16	16
9	10	10
10	17	19
11	135	115
12	133	110
13	33	32
14	40	39
15	44	43
16	30	31
17	48	49
18	78	80
19	16	14
20	22	21
21	429	408
22	447	426
23	16	15
24	87	86
25	17	16
26	26	18
27	16	13
28	14	14
29	58	48
30	21	21
31	19	16
32	48	37
33	13	10
34	79	64
35	98	80
36	141	113
37	217	210
38	988	968

Supplementary Table 5.2. Number of genes in each gene set and the number that mapped to autosomal coordinates.

Index	Genes in Overlap	Proportion Overlap
1	9	0.75
2	8	1.00
3	13	0.59
4	12	0.55
5	23	0.68
6	27	0.84
7	27	0.64
8	11	0.69
9	8	0.80
10	11	0.65
11	81	0.60
12	80	0.60
13	28	0.85
14	27	0.68
15	39	0.89
16	30	1.00
17	48	1.00
18	78	1.00
19	14	0.88
20	17	0.77
21	321	0.75
22	314	0.70
23	14	0.88
24	70	0.80
25	16	0.94
26	18	0.69
27	13	0.81
28	13	0.93
29	40	0.67
30	14	0.67
31	12	0.63
32	32	0.67
33	9	0.69
34	60	0.76
35	76	0.78
36	104	0.74
37	166	0.76
38	968	1.00

Supplementary Table 5.3. Gene sets were selected from the MSigDB that showed equal to or greater than 50% overlap with the list of NEM genes from MitoCarta.

Index	Genomic Regions Containing Genes (European)	Genes in Genomic Regions (European)	Genomic Regions Containing Genes (AA)	Genes in Genomic Regions (AA)
1	3	3	4	4
2	4	4	3	3
3	12	9	12	9
4	12	10	10	8
5	13	13	9	11
6	10	11	10	11
7	14	13	8	9
8	9	7	7	6
9	3	3	1	1
10	7	5	6	4
11	44	44	31	30
12	43	39	34	31
13	14	14	8	8
14	23	22	13	13
15	16	16	8	8
16	13	13	10	9
17	19	18	14	13
18	31	31	23	22
19	4	4	3	3
20	9	10	5	4
21	153	161	113	118
22	156	176	114	121
23	7	7	4	4
24	28	28	23	24
25	8	8	4	4
26	10	10	6	6
27	4	4	3	3
28	4	4	2	2
29	18	18	17	17
30	10	8	7	6
31	7	6	6	6
32	17	17	12	13
33	5	4	5	5
34	24	24	23	23
35	29	29	26	26
36	43	44	36	37
37	91	92	56	60
38	324	378	248	277

Supplementary Table 5.4. Distribution of genes for each gene set mapped into genomic regions for European and African American samples.

Reactome Pyruvate and TCA Cycle	t statistic	p-value	Study	Expression Atlas Reference
ADHFE1	3.2	0.03	Kennedy *et al.* (2010)	E-GEOD-22097
CS	5.8	0.005	Pavlinkova *et al.* (2009)	E-GEOD-41095
DLAT	3.3	0.036	Pavlinkova *et al.* (2009)	E-GEOD-41095
DLD	3.9	0.016		
NNT	5	0.017	Pavlinkova *et al.* (2009)	E-MEXP-2559
IDH2	7	$3.4x10^{-4}$	Kennedy *et al.* (2010)	E-GEOD-22097
LDHB	-	-	-	-
MDH2	-	-	-	-
ACO2	-	-	-	-
PDK1	8.5	$9.1x10^{-5}$	Kennedy *et al.* (2010)	E-GEOD-22097
SDHC	4.2	0.04	Vukkadapu *et al.* (2005)	E-GEOD-1623
SDHD	5.1	0.007	Pavlinkova *et al.* (2009)	E-GEOD-41095
BSG	4.9	0.02	Vukkadapu *et al.* (2005)	E-GEOD-1623
L2HGDH	3.5	0.029	Pavlinkova *et al.* (2009)	E-GEOD-41095
PDHX	5.4	0.001	Gerber *et al.* (2006)	E-GEOD-4745
SUCLG2	4.2	0.015	Pavlinkova *et al.* (2009)	E-GEOD-41095

Supplementary Table 5.5. Differential expression studies in diabetes case-control studies from the Expression Atlas for genes of the Reactome pyruvate and TCA cycle.

KEGG TCA Cycle	t statistic	p-value	Study	Expression Atlas Reference
CS	5.8	0.005	Pavlinkova *et al.* (2009)	E-GEOD-41095
DLAT	3.3	0.036	Pavlinkova *et al.* (2009)	E-GEOD-41095
DLD	3.9	0.016	Mishra *et al.* (2004)	E-GEOD-642
IDH2	7	$3.4x10^{-4}$	Kennedy *et al.* (2010)	E-GEOD-22097
MDH1	3.3	0.029	Kennedy *et al.* (2010)	E-GEOD-22097
MDH2	2.3	0.704	van Tienen et al. (2011)*	E-GEOD-19420
ACO2	-	-	-	-
SDHC	4.2	0.04	Vukkadapu *et al.* (2005)	E-GEOD-1623
SDHD	5.1	0.007	Pavlinkova *et al.* (2009)	E-GEOD-41095
SUCLG2	4.2	0.015	Pavlinkova *et al.* (2009)	E-GEOD-41095
SUCLG1	-	-	-	-

Supplementary Table 5.6. Differential expression studies in diabetes case-control studies from the Expression Atlas for genes of the KEGG TCA cycle.
*No bibliographic information available.

MIPS 28S Ribosomal Subunit	t statistic	p-value	Study	Expression Atlas Reference
DAP3	4.8	0.004	Kennedy *et al.* (2010)	E-GEOD-22097
MRPS11	-	-	-	-
MRPS16	6.6	0.003	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS21	7.4	0.002	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS22	6.3	$7.1x10^{-4}$	Kennedy *et al.* (2010)	E-GEOD-22097
MRPS23	3.6	0.026	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS26	3	0.047	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS27	7.1	0.002	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS28	5.6	0.005	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS31	5.9	0.004	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS33	5.1	0.003	Pavlinkova *et al.* (2009)	E-GEOD-22097
MRPS34	4	0.017	Pavlinkova *et al.* (2009)	E-GEOD-41095
MRPS35	3.2	0.033	Kennedy *et al.* (2010)	E-GEOD-22097

Supplementary Table 5.7. Differential expression studies in diabetes case-control studies from the Expression Atlas for genes of the MIPS 28S ribosomal subunit.

Bibliography

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

Alibes, A., Yankilevich, P., Canada, A. & Diaz-Uriarte, R. 2007. IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics,* 8**,** 9.

Allen, J. F. 2003. The function of genomes in bioenergetic organelles. *Philos Trans R Soc Lond B Biol Sci,* 358**,** 19-37; discussion 37-8.

American Diabetes, A. 2011. Diagnosis and classification of diabetes mellitus. *Diabetes Care,* 34 Suppl 1**,** S62-9.

Andrew, T., Aviv, A., Falchi, M., Surdulescu, G. L., Gardner, J. P., Lu, X., Kimura, M., Kato, B. S., Valdes, A. M. & Spector, T. D. 2006. Mapping genetic loci that determine leukocyte telomere length in a large sample of unselected female sibling pairs. *Am J Hum Genet,* 78**,** 480-6.

Andrew, T., Hart, D. J., Snieder, H., de Lange, M., Spector, T. D. & MacGregor, A. J. 2001. Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin Res,* 4**,** 464-77.

Aragon, G. & Younossi, Z. M. 2010. When and how to evaluate mildly elevated liver enzymes in apparently healthy patients. *Cleve Clin J Med,* 77**,** 195-204.

Armstrong, M. J., Houlihan, D. D., Bentham, L., Shaw, J. C., Cramb, R., Olliff, S., Gill, P. S., Neuberger, J. M., Lilford, R. J. & Newsome, P. N. 2012. Presence and severity of non-alcoholic fatty liver disease in a large prospective primary care cohort. *Journal of Hepatology,* 56**,** 234-240.

Ashcroft, F. M. & Rorsman, P. 2012. Diabetes mellitus and the beta cell: the last ten years. *Cell,* 148**,** 1160-71.

Askland, K., Read, C. & Moore, J. 2009. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum Genet,* 125**,** 63-79.

Bastard, J. P., Antuna-Puente, B., Feve, B. & Fellahi, S. 2008. Adipokines: The missing link between insulin resistance and obesity. *Diabetes & Metabolism,* 34**,** 2-11.

Baughman, J. M., Nilsson, R., Gohil, V. M., Arlow, D. H., Gauhar, Z. & Mootha, V. K. 2009. A computational screen for regulators of oxidative phosphorylation implicates SLIRP in mitochondrial RNA homeostasis. *PLoS Genet,* 5**,** e1000590.

Benzaquen, B. S. & Nguyen-Thanh, H. T. 2009. Screening for Subclinical Coronary Artery Disease Measuring Carotid Intima Media Thickness. *American Journal of Cardiology,* 104**,** 1383-1388.

Berg, J., Hung, Y. P. & Yellen, G. 2009. A genetically encoded fluorescent reporter of ATP:ADP ratio. *Nat Methods,* 6**,** 161-6.

Bertin, E., Marcus, C., Ruiz, J. C., Eschard, J. P. & Leutenegger, M. 2000. Measurement of visceral adipose tissue by DXA combined with anthropometry in obese humans. *International Journal of Obesity,* 24**,** 263-270.

Bodmer, W. & Bonilla, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet,* 40**,** 695-701.

Bonnycastle, L. L., Willer, C. J., Conneely, K. N., Jackson, A. U., Burrill, C. P., Watanabe, R. M., Chines, P. S., Narisu, N., Scott, L. J., Enloe, S. T., *et al.* 2006. Common variants in maturity-onset diabetes of the young genes contribute to risk of type 2 diabetes in Finns. *Diabetes,* 55**,** 2534-40.

Bowman, P., Flanagan, S. E., Edghill, E. L., Damhuis, A., Shepherd, M. H., Paisey, R., Hattersley, A. T. & Ellard, S. 2012. Heterozygous ABCC8 mutations are a cause of MODY. *Diabetologia,* 55**,** 123-7.

Brown, M. S. & Goldstein, J. L. 2008. Selective versus total insulin resistance: A pathogenic paradox. *Cell Metabolism,* 7**,** 95-96.

Buchanan, C. C., Torstenson, E. S., Bush, W. S. & Ritchie, M. D. 2012. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *J Am Med Inform Assoc,* 19**,** 289-94.

Buchner, D. A., Yazbek, S. N., Solinas, P., Burrage, L. C., Morgan, M. G., Hoppel, C. L. & Nadeau, J. H. 2011. Increased mitochondrial oxidative phosphorylation in the liver is associated with obesity and insulin resistance. *Obesity (Silver Spring),* 19**,** 917-24.

Bugger, H., Chen, D., Riehle, C., Soto, J., Theobald, H. A., Hu, X. X., Ganesan, B., Weimer, B. C. & Abel, E. D. 2009. Tissue-specific remodeling of the mitochondrial proteome in type 1 diabetic akita mice. *Diabetes,* 58**,** 1986-97.

Burke, M. D. 2002. Liver function: test selection and interpretation of results. *Clin Lab Med,* 22**,** 377-90.

Calvo, S., Jain, M., Xie, X., Sheth, S. A., Chang, B., Goldberger, O. A., Spinazzola, A., Zeviani, M., Carr, S. A. & Mootha, V. K. 2006. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet,* 38**,** 576-82.

Calvo, S. E. & Mootha, V. K. 2010. The mitochondrial proteome and human disease. *Annu Rev Genomics Hum Genet,* 11**,** 25-44.

Cecelja, M., Hussain, M., Greil, G., Spector, T. D. & Chowienczyk, P. 2010. Arterial stiffening relates to arterial calcification but not to non-calcified atheroma as determined by multimodality imaging of the aorta. *Journal of Human Hypertension,* 24**,** 693-693.

Chan, D. C. 2006. Mitochondria: dynamic organelles in disease, aging, and development. *Cell,* 125**,** 1241-52.

Chinopoulos, C., Vajda, S., Csanady, L., Mandi, M., Mathe, K. & Adam-Vizi, V. 2009. A novel kinetic assay of mitochondrial ATP-ADP exchange rate mediated by the ANT. *Biophys J,* 96**,** 2490-504.

Chomentowski, P., Coen, P. M., Radikova, Z., Goodpaster, B. H. & Toledo, F. G. 2011. Skeletal muscle mitochondria in insulin resistance: differences in intermyofibrillar versus subsarcolemmal subpopulations and relationship to metabolic flexibility. *J Clin Endocrinol Metab,* 96**,** 494-503.

Choo, H. J., Kim, J. H., Kwon, O. B., Lee, C. S., Mun, J. Y., Han, S. S., Yoon, Y. S., Yoon, G., Choi, K. M. & Ko, Y. G. 2006. Mitochondria are impaired in the adipocytes of type 2 diabetic mice. *Diabetologia,* 49**,** 784-91.

Civitarese, A. E., MacLean, P. S., Carling, S., Kerr-Bayles, L., McMillan, R. P., Pierce, A., Becker, T. C., Moro, C., Finlayson, J., Lefort, N., *et al.* 2010. Regulation of skeletal muscle oxidative capacity and insulin signaling by the mitochondrial rhomboid protease PARL. *Cell Metab,* 11**,** 412-26.

Claros, M. G. & Vincens, P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem,* 241**,** 779-86.

Clayton, D. G. 2009. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet,* 5**,** e1000540.

Collins, A. & Morton, N. E. 1998. Mapping a disease locus by allelic association. *Proc Natl Acad Sci U S A,* 95**,** 1741-5.

Curran, J. E., Jowett, J. B., Abraham, L. J., Diepeveen, L. A., Elliott, K. S., Dyer, T. D., Kerr-Bayles, L. J., Johnson, M. P., Comuzzie, A. G., Moses, E. K., *et al.* 2010. Genetic variation in PARL influences mitochondrial content. *Hum Genet,* 127**,** 183-90.

Danaei, G., Finucane, M. M., Lu, Y., Singh, G. M., Cowan, M. J., Paciorek, C. J., Lin, J. K., Farzadfar, F., Khang, Y. H., Stevens, G. A., *et al.* 2011. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet,* 378**,** 31-40.

David Clayton, M. H. 1993. *Statistical Models in Epidemiology,* Oxford University Press.

De Lucia Rolfe, E., Norris, S. A., Sleigh, A., Brage, S., Dunger, D. B., Stolk, R. P. & Ong, K. K. 2011. Validation of ultrasound estimates of visceral fat in black South African adolescents. *Obesity (Silver Spring),* 19**,** 1892-7.

Dean, M. & Allikmets, R. 2001. Complete characterization of the human ABC gene family. *J Bioenerg Biomembr,* 33**,** 475-9.

DeFronzo, R. A., Tobin, J. D. & Andres, R. 1979. Glucose clamp technique: a method for quantifying insulin secretion and resistance. *Am J Physiol,* 237**,** E214-23.

DeWan, A., Klein, R. J. & Hoh, J. 2007. Linkage disequilibrium mapping for complex disease genes. *Methods Mol Biol,* 376**,** 85-107.

Diabetes Genetics Initiative of Broad Institute of, H., Mit, L. U., Novartis Institutes of BioMedical, R., Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., *et al.* 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science,* 316**,** 1331-6.

Dresner, A., Laurent, D., Marcucci, M., Griffin, M. E., Dufour, S., Cline, G. W., Slezak, L. A., Andersen, D. K., Hundal, R. S., Rothman, D. L., *et al.* 1999. Effects of free fatty acids on glucose transport and IRS-1-associated phosphatidylinositol 3-kinase activity. *J Clin Invest,* 103**,** 253-9.

Dudbridge, F. & Gusnanto, A. 2008. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol,* 32**,** 227-34.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet,* 11**,** 446-50.

Elbers, C. C., van Eijk, K. R., Franke, L., Mulder, F., van der Schouw, Y. T., Wijmenga, C. & Onland-Moret, N. C. 2009. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol,* 33**,** 419-31.

Elding, H., Lau, W., Swallow, D. M. & Maniatis, N. 2011. Dissecting the genetics of complex inheritance: linkage disequilibrium mapping provides insight into Crohn disease. *Am J Hum Genet,* 89**,** 798-805.

Elding, H., Lau, W., Swallow, D. M. & Maniatis, N. 2013. Refinement in localization and identification of gene regions associated with Crohn disease. *Am J Hum Genet,* 92**,** 107-13.

Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol,* 300**,** 1005-16.

Escobar Galvis, M. L., Allen, J. F. & Hakansson, G. 1998. Protein synthesis by isolated pea mitochondria is dependent on the activity of respiratory complex II. *Curr Genet,* 33**,** 320-9.

Evans, D. M., Gillespie, N. A. & Martin, N. G. 2002. Biometrical genetics. *Biol Psychol,* 61**,** 33-51.

Fabbrini, E., Magkos, F., Mohammed, B. S., Pietka, T., Abumrad, N. A., Patterson, B. W., Okunade, A. & Klein, S. 2009. Intrahepatic fat, not visceral fat, is linked with metabolic complications of obesity. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 15430-15435.

Falconer, D. S. 1989. *Introduction to quantitative genetics,* Burnt Mill, Harlow, Essex, England

New York, Longman

Wiley.

Fawcett, K. A., Wareham, N. J., Luan, J., Syddall, H., Cooper, C., O'Rahilly, S., Day, I. N., Sandhu, M. S. & Barroso, I. 2006. PARL Leu262Val is not associated with fasting insulin levels in UK populations. *Diabetologia,* 49**,** 2649-52.

Fisher, R. 1925. *Statistical Methods for Research Workers,* Oliver and Boyd (Edinburgh).

Fisk, N. M., Duncombe, G. J. & Sullivan, M. H. 2009. The basic and clinical science of twin-twin transfusion syndrome. *Placenta,* 30**,** 379-90.

Florez, J. C., Burtt, N., de Bakker, P. I., Almgren, P., Tuomi, T., Holmkvist, J., Gaudet, D., Hudson, T. J., Schaffner, S. F., Daly, M. J., *et al.* 2004. Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes,* 53**,** 1360-8.

Foretz, M., Hebrard, S., Leclerc, J., Zarrinpashneh, E., Soty, M., Mithieux, G., Sakamoto, K., Andreelli, F. & Viollet, B. 2010. Metformin inhibits hepatic gluconeogenesis in mice independently of the LKB1/AMPK pathway via a decrease in hepatic energy state. *J Clin Invest,* 120**,** 2355-69.

Fox, C. S., Massaro, J. M., Hoffmann, U., Pou, K. M., Maurovich-Horvat, P., Liu, C. Y., Vasan, R. S., Murabito, J. M., Meigs, J. B., Cupples, L. A., *et al.* 2007. Abdominal visceral and subcutaneous adipose tissue compartments - Association with metabolic risk factors in the Framingham Heart Study. *Circulation,* 116**,** 39-48.

Fradkin, J. E. & Rodgers, G. P. 2013. Diabetes research: a perspective from the National Institute of Diabetes and Digestive and Kidney Diseases. *Diabetes,* 62**,** 320-6.

Francke, S., Manraj, M., Lacquemant, C., Lecoeur, C., Lepretre, F., Passa, P., Hebe, A., Corset, L., Yan, S. L., Lahmidi, S., *et al.* 2001. A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. *Hum Mol Genet,* 10**,** 2751-65.

Frayling, T. M. 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet,* 8**,** 657-62.

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I. & Wright, F. A. 2010. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics,* 11**,** 574.

Gene Ontology, C. 2013. Gene Ontology annotations and resources. *Nucleic Acids Res,* 41**,** D530-5.

George, S., Rochford, J. J., Wolfrum, C., Gray, S. L., Schinner, S., Wilson, J. C., Soos, M. A., Murgatroyd, P. R., Williams, R. M., Acerini, C. L., *et al.* 2004. A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science,* 304**,** 1325-8.

Gerber, L. K., Aronow, B. J. & Matlib, M. A. 2006. Activation of a novel long-chain free fatty acid generation and export system in mitochondria of diabetic rat hearts. *Am J Physiol Cell Physiol,* 291**,** C1198-207.

Giral, P., Ratziu, V., Couvert, P., Carrie, A., Kontush, A., Girerd, X. & Chapman, M. J. 2010. Plasma bilirubin and gamma-glutamyltransferase activity are inversely related in dyslipidemic patients with metabolic syndrome: Relevance to oxidative stress. *Atherosclerosis,* 210**,** 607-613.

Gloyn, A. L., Weedon, M. N., Owen, K. R., Turner, M. J., Knight, B. A., Hitman, G., Walker, M., Levy, J. C., Sampson, M., Halford, S., *et al.* 2003. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes,* 52**,** 568-72.

Goeman, J. J. & Buhlmann, P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics,* 23**,** 980-7.

Greeley, S. A., Naylor, R. N., Philipson, L. H. & Bell, G. I. 2011. Neonatal diabetes: an expanding list of genes allows for improved diagnosis and treatment. *Curr Diab Rep,* 11**,** 519-32.

Guan, W., Pluzhnikov, A., Cox, N. J., Boehnke, M. & International Type 2 Diabetes Linkage Analysis, C. 2008. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Hum Hered,* 66**,** 35-49.

Guda, C., Guda, P., Fahy, E. & Subramaniam, S. 2004. MITOPRED: a web server for the prediction of mitochondrial proteins. *Nucleic Acids Res,* 32**,** W372-4.

Halimi, S. 2006. Metformin: 50 years old, fit as a fiddle, and indispensable for its pivotal role in type 2 diabetes management. *Diabetes Metab,* 32**,** 555-6.

Hancock, C. R., Han, D. H., Chen, M., Terada, S., Yasuda, T., Wright, D. C. & Holloszy, J. O. 2008. High-fat diets cause insulin resistance despite an increase in muscle mitochondria. *Proc Natl Acad Sci U S A,* 105**,** 7815-20.

Hani, E. H., Stoffers, D. A., Chevre, J. C., Durand, E., Stanojevic, V., Dina, C., Habener, J. F. & Froguel, P. 1999. Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus. *J Clin Invest,* 104**,** R41-8.

Hart, L. M., Hansen, T., Rietveld, I., Dekker, J. M., Nijpels, G., Janssen, G. M., Arp, P. A., Uitterlinden, A. G., Jorgensen, T., Borch-Johnsen, K., *et al.* 2005. Evidence that the mitochondrial leucyl tRNA synthetase (LARS2) gene represents a novel type 2 diabetes susceptibility gene. *Diabetes,* 54**,** 1892-5.

Hatunic, M., Stapleton, M., Hand, E., DeLong, C., Crowley, V. E. & Nolan, J. J. 2009. The Leu262Val polymorphism of presenilin associated rhomboid like protein (PARL) is associated with earlier onset of type 2 diabetes and increased urinary microalbumin creatinine ratio in an Irish case-control population. *Diabetes Res Clin Pract,* 83**,** 316-9.

Hayashi, T., Boyko, E. J., Leonetti, D. L., McNeely, M. J., Newell-Morris, L., Kahn, S. E. & Fujimoto, W. Y. 2004. Visceral adiposity is an independent predictor of incident hypertension in Japanese Americans. *Annals of Internal Medicine,* 140**,** 992-1000.

Hedrick, P. W. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics,* 117**,** 331-41.

Hegele, R. A., Sun, F., Harris, S. B., Anderson, C., Hanley, A. J. & Zinman, B. 1999. Genome-wide scanning for type 2 diabetes susceptibility in Canadian Oji-Cree, using 190 microsatellite markers. *J Hum Genet,* 44**,** 10-4.

Henquin, J. C. 2009. Regulation of insulin secretion: a matter of phase control and amplitude modulation. *Diabetologia,* 52**,** 739-51.

Henquin, J. C. 2011. The dual control of insulin secretion by glucose involves triggering and amplifying pathways in beta-cells. *Diabetes Res Clin Pract,* 93 Suppl 1**,** S27-31.

Hill, A. M., LaForgia, J., Coates, A. M., Buckley, J. D. & Howe, P. R. C. 2007. Estimating abdominal adipose tissue with DXA and anthropometry. *Obesity,* 15**,** 504-510.

Hill, W. G. & Robertson, A. 1968. The effects of inbreeding at loci with heterozygote advantage. *Genetics,* 60**,** 615-28.

Himsworth, H. P. 1936. Management of Diabetes Mellitus. *Br Med J,* 2**,** 188-90.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., Wellcome Trust Case-Control, C., Owen, M. J., O'Donovan, M. C. &

Craddock, N. 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet,* 85**,** 13-24.

Hong, M. G., Pawitan, Y., Magnusson, P. K. & Prince, J. A. 2009. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum Genet,* 126**,** 289-301.

Hong, Y. L., Rice, T., Gagnon, J., Despres, J. P., Nadeau, A., Perusse, L., Bouchard, C., Leon, A. S., Skinner, J. S., Wilmore, J. H., *et al.* 1998. Familial clustering of insulin and abdominal visceral fat: The HERITAGE family study. *Journal of Clinical Endocrinology & Metabolism,* 83**,** 4239-4245.

Huang da, W., Sherman, B. T. & Lempicki, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

Huang da, W., Sherman, B. T., Stephens, R., Baseler, M. W., Lane, H. C. & Lempicki, R. A. 2008. DAVID gene ID conversion tool. *Bioinformation,* 2**,** 428-30.

Imamura, H., Nhat, K. P., Togawa, H., Saito, K., Iino, R., Kato-Yamada, Y., Nagai, T. & Noji, H. 2009. Visualization of ATP levels inside single living cells with fluorescence resonance energy transfer-based genetically encoded indicators. *Proc Natl Acad Sci U S A,* 106**,** 15651-6.

Irlbeck, T., Massaro, J. M., Bamberg, F., O'Donnell, C. J., Hoffmann, U. & Fox, C. S. 2010. Association between single-slice measurements of visceral and abdominal subcutaneous adipose tissue with volumetric measurements: the Framingham Heart Study. *International Journal of Obesity,* 34**,** 781-787.

Jacobs, H. T. & Turnbull, D. M. 2005. Nuclear genes and mitochondrial translation: a new class of genetic disease. *Trends Genet,* 21**,** 312-4.

Kahn, C. R., Flier, J. S., Bar, R. S., Archer, J. A., Gorden, P., Martin, M. M. & Roth, J. 1976. The syndromes of insulin resistance and acanthosis nigricans. Insulin-receptor disorders in man. *N Engl J Med,* 294**,** 739-45.

Kanehisa, M. & Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

Karelis, A. D., Rabasa-Lhoret, R., Pompilus, R., Messier, V., Strychar, I., Brochu, M. & Aubertin-Leheudre, M. 2012. Relationship between the Bertin index to estimate visceral adipose tissue from dual-energy X-ray absorptiometry and cardiometabolic risk factors before and after weight loss. *Obesity (Silver Spring),* 20**,** 886-90.

Kariv, R., Leshno, M., Beth-Or, A., Strul, H., Blendis, L., Kokia, E., Noff, D., Zelber-Sagie, S., Sheinberg, B., Oren, R., *et al.* 2006. Re-evaluation of serum alanine aminotransferase upper normal limit and its modulating factors in a large-scale population study. *Liver International,* 26**,** 445-450.

Karpe, F., Dickmann, J. R. & Frayn, K. N. 2011. Fatty acids, obesity, and insulin resistance: time for a reevaluation. *Diabetes,* 60**,** 2441-9.

Keeling, P. J. & Palmer, J. D. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet,* 9**,** 605-18.

Kelley, D. E., He, J., Menshikova, E. V. & Ritov, V. B. 2002. Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. *Diabetes,* 51**,** 2944-50.

Kennedy, J., Katsuta, H., Jung, M. H., Marselli, L., Goldfine, A. B., Balis, U. J., Sgroi, D., Bonner-Weir, S. & Weir, G. C. 2010. Protective unfolded protein response in human pancreatic beta cells transplanted into mice. *PLoS One,* 5**,** e11211.

Khatri, P. & Draghici, S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics,* 21**,** 3587-95.

Kim, J. Y., De Wall, E. V., Laplante, M., Azzara, A., Trujillo, M. E., Hofmann, S. M., Schraw, T., Durand, J. L., Li, H., Li, G., *et al.* 2007. Obesity-associated improvements in metabolic profile through expansion of adipose tissue. *Journal of Clinical Investigation,* 117**,** 2621-2637.

Kim, S. K., Park, S. W., Kim, S. H., Cha, B. S., Lee, H. C. & Cho, Y. W. 2009. Visceral fat amount is associated with carotid atherosclerosis even in type 2 diabetic men with a normal waist circumference. *International Journal of Obesity,* 33**,** 131-135.

Kissebah, A. H., Sonnenberg, G. E., Myklebust, J., Goldstein, M., Broman, K., James, R. G., Marks, J. A., Krakower, G. R., Jacob, H. J., Weber, J., *et al.* 2000. Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. *Proc Natl Acad Sci U S A,* 97**,** 14478-83.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., *et al.* 2005. Complement factor H polymorphism in age-related macular degeneration. *Science,* 308**,** 385-9.

Koves, T. R., Ussher, J. R., Noland, R. C., Slentz, D., Mosedale, M., Ilkayeva, O., Bain, J., Stevens, R., Dyck, J. R., Newgard, C. B., *et al.* 2008. Mitochondrial overload and incomplete fatty acid oxidation contribute to skeletal muscle insulin resistance. *Cell Metab,* 7**,** 45-56.

Kramer, C. K., von Muhlen, D., Gross, J. L. & Barrett-Connor, E. 2009. A Prospective Study of Abdominal Obesity and Coronary Artery Calcium Progression in Older Adults. *Journal of Clinical Endocrinology & Metabolism,* 94**,** 5039-5044.

Kumar, A. 2009. An overview of nested genes in eukaryotic genomes. *Eukaryot Cell,* 8**,** 1321-9.

Kuo, T. Y., Lau, W. & Collins, A. R. 2007. LDMAP: the construction of high-resolution linkage disequilibrium maps of the human genome. *Methods Mol Biol,* 376**,** 47-57.

Kusminski, C. M. & Scherer, P. E. 2012. Mitochondrial dysfunction in white adipose tissue. *Trends Endocrinol Metab,* 23**,** 435-43.

Laye, M. J., Rector, R. S., Warner, S. O., Naples, S. P., Perretta, A. L., Uptergrove, G. M., Laughlin, M. H., Thyfault, J. P., Booth, F. W. & Ibdah, J. A. 2009. Changes in visceral adipose tissue mitochondrial content with type 2 diabetes and daily voluntary wheel running in OLETF rats. *J Physiol,* 587**,** 3729-39.

Lazo, M., Selvin, E. & Clark, J. M. 2008. Brief communication: Clinical implications of short-term variability in liver function test results. *Annals of Internal Medicine,* 148**,** 348-W76.

Lear, S. A., Humphries, K. H., Kohli, S., Frohlich, J. J., Birmingham, C. L. & Mancini, G. B. J. 2007. Visceral adipose tissue, a potential risk factor for carotid atherosclerosis - Results of the multicultural community health assessment trial (M-CHAT). *Stroke,* 38**,** 2422-2429.

Lehninger, A. L. 1979. *Biochemistry. The Molecular Basis of Cell Structure and Function.,* New York, Worth Publishers, Inc.

Leister, D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet,* 21**,** 655-63.

Leney, S. E. & Tavare, J. M. 2009. The molecular basis of insulin-stimulated glucose uptake: signalling, trafficking and potential drug targets. *J Endocrinol,* 203**,** 1-18.

Leslie, R. D. 2010. Predicting Adult-Onset Automunune Diabetes Clarity From Complexity. *Diabetes,* 59**,** 330-331.

Lettre, G., Lange, C. & Hirschhorn, J. N. 2007. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol,* 31**,** 358-62.

Levy, E., Lalonde, G., Delvin, E., Elchebly, M., Precourt, L. P., Seidah, N. G., Spahis, S., Rabasa-Lhoret, R. & Ziv, E. 2010. Intestinal and hepatic cholesterol carriers in diabetic Psammomys obesus. *Endocrinology,* 151**,** 958-70.

Levy, J. C., Matthews, D. R. & Hermans, M. P. 1998. Correct homeostasis model assessment (HOMA) evaluation uses the computer program. *Diabetes Care,* 21**,** 2191-2.

Lewis, C. M. & Knight, J. 2012. Introduction to genetic association studies. *Cold Spring Harb Protoc,* 2012**,** 297-306.

Lewontin, R. C. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics,* 49**,** 49-67.

Liesa, M., Palacin, M. & Zorzano, A. 2009. Mitochondrial dynamics in mammalian health and disease. *Physiol Rev,* 89**,** 799-845.

Lin, L. Y., Kuo, H. K., Hwang, J. J., Lai, L. P., Chiang, F. T., Tseng, C. D. & Lin, J. L. 2009. Serum bilirubin is inversely associated with insulin resistance and metabolic syndrome among children and adolescents. *Atherosclerosis,* 203**,** 563-568.

Lorenzo, C., Haffner, S. M., Stancakova, A. & Laakso, M. 2010. Relation of direct and surrogate measures of insulin resistance to cardiovascular risk factors in nondiabetic finnish offspring of type 2 diabetic individuals. *J Clin Endocrinol Metab,* 95**,** 5082-90.

Love-Gregory, L. D., Wasson, J., Ma, J., Jin, C. H., Glaser, B., Suarez, B. K. & Permutt, M. A. 2004. A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor-4 alpha gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an ashkenazi jewish population. *Diabetes,* 53**,** 1134-40.

Maassen, J. A., LM, T. H., Van Essen, E., Heine, R. J., Nijpels, G., Jahangir Tafrechi, R. S., Raap, A. K., Janssen, G. M. & Lemkes, H. H. 2004. Mitochondrial diabetes: molecular mechanisms and clinical presentation. *Diabetes,* 53 Suppl 1**,** S103-9.

Maechler, P., Carobbio, S. & Rubi, B. 2006. In beta-cells, mitochondria integrate and generate metabolic signals controlling insulin secretion. *Int J Biochem Cell Biol,* 38**,** 696-709.

Maes, H. H., Neale, M. C. & Eaves, L. J. 1997. Genetic and environmental factors in relative body weight and human adiposity. *Behav Genet,* 27**,** 325-51.

Malécot, G. 1970. *The mathematics of heredity,* San Francisco,, W. H. Freeman.

Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W. & Morton, N. E. 2004. Positional cloning by linkage disequilibrium. *Am J Hum Genet,* 74**,** 846-55.

Maniatis, N., Collins, A., Xu, C. F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. 2002. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A,* 99**,** 2228-33.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al.* 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

Mantzoros, C. S., Fiorenza, C. G. & Chou, S. H. 2011. Lipodystrophy: pathophysiology and advances in treatment. *Nature Reviews Endocrinology,* 7**,** 137-150.

Mantzoros, C. S., Liolios, A. D., Tritos, N. A., Kaklamani, V. G., Doulgerakis, D. E., Griveas, I., Moses, A. C. & Flier, J. S. 1998. Circulating insulin

concentrations, smoking, and alcohol intake are important independent predictors of leptin in young healthy men. *Obes Res,* 6**,** 179-86.

Martin, N., Boomsma, D. & Machin, G. 1997. A twin-pronged attack on complex traits. *Nature Genetics,* 17**,** 387-392.

Maston, G. A., Evans, S. K. & Green, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

Matsuo, M. 2010. ATP-binding cassette proteins involved in glucose and lipid homeostasis. *Biosci Biotechnol Biochem,* 74**,** 899-907.

Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F. & Turner, R. C. 1985. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia,* 28**,** 412-9.

McLaughlin, T., Abbasi, F., Lamendola, C. & Reaven, G. 2007. Heterogeneity in the prevalence of risk factors for cardiovascular disease and type 2 diabetes mellitus in obese individuals: effect of differences in insulin sensitivity. *Arch Intern Med,* 167**,** 642-8.

Mercer, T. R., Neph, S., Dinger, M. E., Crawford, J., Smith, M. A., Shearwood, A. M., Haugen, E., Bracken, C. P., Rackham, O., Stamatoyannopoulos, J. A., *et al.* 2011. The human mitochondrial transcriptome. *Cell,* 146**,** 645-58.

Mishra, R., Emancipator, S. N., Miller, C., Kern, T. & Simonson, M. S. 2004. Adipose differentiation-related protein and regulators of lipid homeostasis identified by gene expression profiling in the murine db/db diabetic kidney. *Am J Physiol Renal Physiol,* 286**,** F913-21.

Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. 2013. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res Hum Genet,* 16**,** 144-9.

Moayyeri, A., Hammond, C. J., Valdes, A. M. & Spector, T. D. 2012. Cohort Profile: TwinsUK and Healthy Ageing Twin Study. *Int J Epidemiol*.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet,* 34**,** 267-73.

Mori, Y., Otabe, S., Dina, C., Yasuda, K., Populaire, C., Lecoeur, C., Vatin, V., Durand, E., Hara, K., Okada, T., *et al.* 2002. Genome-wide search for type 2 diabetes in Japanese affected sib-pairs confirms susceptibility genes on 3q, 15q, and 20q and identifies two new candidate Loci on 7p and 11p. *Diabetes,* 51**,** 1247-55.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., *et al.* 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet,* 44**,** 981-90.

Muniyappa, R., Iantorno, M. & Quon, M. J. 2008. An integrated view of insulin resistance and endothelial dysfunction. *Endocrinol Metab Clin North Am,* 37**,** 685-711, ix-x.

Muoio, D. M. & Newgard, C. B. 2008. Mechanisms of disease: molecular and metabolic mechanisms of insulin resistance and beta-cell failure in type 2 diabetes. *Nat Rev Mol Cell Biol,* 9**,** 193-205.

Natali, A., Gastaldelli, A., Camastra, S., Sironi, A. M., Toschi, E., Masoni, A., Ferrannini, E. & Mari, A. 2000. Dose-response characteristics of insulin action on glucose metabolism: a non-steady-state approach. *Am J Physiol Endocrinol Metab,* 278**,** E794-801.

Neale, M. C., Cardon, L. R. & North Atlantic Treaty Organization. Scientific Affairs Division. 1992. *Methodology for genetic studies of twins and families,* Dordrecht ; Boston, Kluwer Academic Publishers.

Nenquin, M., Szollosi, A., Aguilar-Bryan, L., Bryan, J. & Henquin, J. C. 2004. Both triggering and amplifying pathways contribute to fuel-induced insulin secretion in the absence of sulfonylurea receptor-1 in pancreatic beta-cells. *J Biol Chem,* 279**,** 32316-24.

Neupert, W. & Herrmann, J. M. 2007. Translocation of proteins into mitochondria. *Annu Rev Biochem,* 76**,** 723-49.

Neve, B., Fernandez-Zapico, M. E., Ashkenazi-Katalan, V., Dina, C., Hamid, Y. H., Joly, E., Vaillant, E., Benmezroua, Y., Durand, E., Bakaher, N., *et al.* 2005. Role of transcription factor KLF11 and its diabetes-associated gene variants in pancreatic beta cell function. *Proc Natl Acad Sci U S A,* 102**,** 4807-12.

Newsholme, P., Gaudel, C. & Krause, M. 2012. Mitochondria and diabetes. An intriguing pathogenetic role. *Adv Exp Med Biol,* 942**,** 235-47.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., *et al.* 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet,* 7**,** e1002003.

Nielsen, S., Guo, Z., Johnson, C. M., Hensrud, D. D. & Jensen, M. D. 2004. Splanchnic lipolysis in human obesity. *J Clin Invest,* 113**,** 1582-8.

Nowicki, M. T., Aleksunes, L. M., Sawant, S. P., Dnyanmote, A. V., Mehendale, H. M. & Manautou, J. E. 2008. Renal and hepatic transporter expression in type 2 diabetic rats. *Drug Metab Lett,* 2**,** 11-7.

O'Brien, R. M. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity,* 41**,** 673-690.

O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D., International Schizophrenia, C. & Corvin, A. 2011. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol Psychiatry,* 16**,** 286-92.

Olsson, A. H., Ronn, T., Ladenvall, C., Parikh, H., Isomaa, B., Groop, L. & Ling, C. 2011a. Two common genetic variants near nuclear-encoded OXPHOS genes are associated with insulin secretion in vivo. *Eur J Endocrinol,* 164**,** 765-71.

Olsson, A. H., Yang, B. T., Hall, E., Taneera, J., Salehi, A., Nitert, M. D. & Ling, C. 2011b. Decreased expression of genes involved in oxidative phosphorylation in human pancreatic islets from patients with type 2 diabetes. *Eur J Endocrinol,* 165**,** 589-95.

Owen, M. R., Doran, E. & Halestrap, A. P. 2000. Evidence that metformin exerts its anti-diabetic effects through inhibition of complex 1 of the mitochondrial respiratory chain. *Biochem J,* 348 Pt 3**,** 607-14.

Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., *et al.* 2008. A mitochondrial protein compendium elucidates complex I disease biology. *Cell,* 134**,** 112-23.

Palmer, L. J. 2007. UK Biobank: bank on it. *Lancet,* 369**,** 1980-2.

Palmer, L. J. & Cardon, L. R. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet,* 366**,** 1223-34.

Palmer, N. D., McDonough, C. W., Hicks, P. J., Roh, B. H., Wing, M. R., An, S. S., Hester, J. M., Cooke, J. N., Bostrom, M. A., Rudock, M. E., *et al.* 2012. A genome-wide association search for type 2 diabetes genes in African Americans. *PLoS One,* 7**,** e29202.

Palou, A., Palou, M., Priego, T., Sanchez, J., Rodriguez, A. M. & Pico, C. 2009. Gene Expression Patterns in Visceral and Subcutaneous Adipose Depots in Rats are Linked to Their Morphologic Features. *Cellular Physiology and Biochemistry,* 24**,** 547-556.

Patti, M. E. & Corvera, S. 2010. The role of mitochondria in the pathogenesis of type 2 diabetes. *Endocr Rev,* 31**,** 364-95.

Pavlinkova, G., Salbaum, J. M. & Kappen, C. 2009. Maternal diabetes alters transcriptional programs in the developing embryo. *BMC Genomics,* 10**,** 274.

Pellegrini, L. & Scorrano, L. 2007. A cut short to death: Parl and Opa1 in the regulation of mitochondrial morphology and apoptosis. *Cell Death Differ,* 14**,** 1275-84.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., *et al.* 2010. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet,* 18**,** 111-7.

Perry, J. R., McCarthy, M. I., Hattersley, A. T., Zeggini, E., Wellcome Trust Case Control, C., Weedon, M. N. & Frayling, T. M. 2009. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes,* 58**,** 1463-7.

Perusse, L., Despres, J. P., Lemieux, S., Rice, T., Rao, D. C. & Bouchard, C. 1996. Familial aggregation of abdominal visceral fat level: Results from the Quebec family study. *Metabolism-Clinical and Experimental,* 45**,** 378-382.

Petrie, J. R., Pearson, E. R. & Sutherland, C. 2011. Implications of genome wide association studies for the understanding of type 2 diabetes pathophysiology. *Biochem Pharmacol,* 81**,** 471-7.

Pfanner, N. & Geissler, A. 2001. Versatility of the mitochondrial protein import machinery. *Nat Rev Mol Cell Biol,* 2**,** 339-49.

Phielix, E., Schrauwen-Hinderling, V. B., Mensink, M., Lenaers, E., Meex, R., Hoeks, J., Kooi, M. E., Moonen-Kornips, E., Sels, J. P., Hesselink, M. K., *et al.* 2008. Lower intrinsic ADP-stimulated mitochondrial respiration underlies in vivo mitochondrial dysfunction in muscle of male type 2 diabetic patients. *Diabetes,* 57**,** 2943-9.

Poole, A. C., Thomas, R. E., Yu, S., Vincow, E. S. & Pallanck, L. 2010. The mitochondrial fusion-promoting factor mitofusin is a substrate of the PINK1/parkin pathway. *PLoS One,* 5**,** e10054.

Postic, C., Dentin, R. & Girard, J. 2004. Role of the liver in the control of carbohydrate and lipid homeostasis. *Diabetes Metab,* 30**,** 398-408.

Powell, B. L., Wiltshire, S., Arscott, G., McCaskie, P. A., Hung, J., McQuillan, B. M., Thompson, P. L., Carter, K. W., Palmer, L. J. & Beilby, J. P. 2008. Association of PARL rs3732581 genetic variant with insulin levels, metabolic syndrome and coronary artery disease. *Hum Genet,* 124**,** 263-70.

Pritchard, J. K. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet,* 69**,** 124-37.

Pritchard, J. K. & Cox, N. J. 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet,* 11**,** 2417-23.

Pritchard, J. K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

Rahmioglu, N., Andrew, T., Cherkas, L., Surdulescu, G., Swaminathan, R., Spector, T. & Ahmadi, K. R. 2009. Epidemiology and Genetic Epidemiology of the Liver Function Test Proteins. *Plos One,* 4.

Ramanan, V. K., Shen, L., Moore, J. H. & Saykin, A. J. 2012. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet,* 28**,** 323-32.

Ravier, M. A., Nenquin, M., Miki, T., Seino, S. & Henquin, J. C. 2009. Glucose controls cytosolic Ca2+ and insulin secretion in mouse islets lacking adenosine triphosphate-sensitive K+ channels owing to a knockout of the pore-forming subunit Kir6.2. *Endocrinology,* 150**,** 33-45.

Reich, D. E. & Lander, E. S. 2001. On the allelic spectrum of human disease. *Trends Genet,* 17**,** 502-10.

Reiling, E., Jafar-Mohammadi, B., van 't Riet, E., Weedon, M. N., van Vliet-Ostaptchouk, J. V., Hansen, T., Saxena, R., van Haeften, T. W., Arp, P. A., Das, S., *et al.* 2010. Genetic association analysis of LARS2 with type 2 diabetes. *Diabetologia,* 53**,** 103-10.

Reiling, E., van Vliet-Ostaptchouk, J. V., van 't Riet, E., van Haeften, T. W., Arp, P. A., Hansen, T., Kremer, D., Groenewoud, M. J., van Hove, E. C., Romijn, J. A., *et al.* 2009. Genetic association analysis of 13 nuclear-encoded mitochondrial candidate genes with type II diabetes mellitus: the DAMAGE study. *Eur J Hum Genet,* 17**,** 1056-62.

Rende, R. D., Plomin, R. & Vandenberg, S. G. 1990. Who discovered the twin method? *Behav Genet,* 20**,** 277-85.

Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. 2008. Use and misuse of the gene ontology annotations. *Nat Rev Genet,* 9**,** 509-15.

Rice, T., Despres, J. P., Daw, E. W., Gagnon, J., Borecki, I. B., Perusse, L., Leon, A. S., Skinner, J. S., Wilmore, J. H., Rao, D. C., *et al.* 1997. Familial resemblance for abdominal visceral fat: the HERITAGE family study. *International Journal of Obesity,* 21**,** 1024-1031.

Richards, J. B., Rivadeneira, F., Inouye, M., Pastinen, T. M., Soranzo, N., Wilson, S. G., Andrew, T., Falchi, M., Gwilliam, R., Ahmadi, K. R., *et al.* 2008. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet,* 371**,** 1505-12.

Romero-Corral, A., Lopez-Jimenez, F., Sierra-Johnson, J. & Somers, V. K. 2008. Differentiating between body fat and lean mass-how should we measure obesity? *Nat Clin Pract Endocrinol Metab,* 4**,** 322-3.

Ruige, J. B., Mertens, I. L., Bartholomeeusen, E., Dirinck, E., Ferrannini, E. & Van Gaal, L. F. 2006. Fasting-based estimates of insulin sensitivity in overweight and obesity: a critical appraisal. *Obesity (Silver Spring),* 14**,** 1250-6.

Sabanayagam, C., Shankar, A., Li, J. L., Pollard, C. & Ducatman, A. 2009. Serum gamma-glutamyl transferase level and diabetes mellitus among US adults. *European Journal of Epidemiology,* 24**,** 369-373.

Samaras, K., McElduff, A., Twigg, S. M., Proietto, J., Prins, J. B., Welborn, T. A., Zimmet, P., Chisholm, D. J. & Campbell, L. V. 2006. Insulin levels in insulin resistance: phantom of the metabolic opera? *Med J Aust,* 185**,** 159-61.

Samuel, V. T., Liu, Z. X., Qu, X. Q., Elder, B. D., Bilz, S., Befroy, D., Romanelli, A. J. & Shulman, G. I. 2004. Mechanism of hepatic insulin resistance in non-alcoholic fatty liver disease. *Journal of Biological Chemistry,* 279**,** 32345-32353.

Samuel, V. T., Petersen, K. F. & Shulman, G. I. 2010. Lipid-induced insulin resistance: unravelling the mechanism. *Lancet,* 375**,** 2267-2277.

Sandhu, M. S., Weedon, M. N., Fawcett, K. A., Wasson, J., Debenham, S. L., Daly, A., Lango, H., Frayling, T. M., Neumann, R. J., Sherva, R., *et al.* 2007. Common variants in WFS1 confer risk of type 2 diabetes. *Nat Genet,* 39**,** 951-3.

Sarkar, D. 2008. *Lattice : multivariate data visualization with R,* New York, Springer.

Savage, D. B., Agostini, M., Barroso, I., Gurnell, M., Luan, J., Meirhaeghe, A., Harding, A. H., Ihrke, G., Rajanayagam, O., Soos, M. A., *et al.* 2002. Digenic inheritance of severe insulin resistance in a human pedigree. *Nat Genet,* 31**,** 379-84.

Savage, D. B., Huang-Doran, I., Sleigh, A., Rochford, J. J. & O'Rahilly, S. 2010. Lipodystrophy: metabolic insights from a rare disorder. *Journal of Endocrinology,* 207**,** 245-255.

Scheper, G. C., van der Knaap, M. S. & Proud, C. G. 2007. Translation matters: protein synthesis defects in inherited disease. *Nat Rev Genet,* 8**,** 711-23.

Schrot, R. J., Patel, K. T. & Foulis, P. 2007. Evaluation of Inaccuracies in the Measurement of Glycemia in the Laboratory, by Glucose Meters, and Through Measurement of Hemoglobin A1c. *Clinical Diabetes,* 25**,** 43-49.

Scorrano, L. 2007. Multiple functions of mitochondria-shaping proteins. *Novartis Found Symp,* 287**,** 47-55; discussion 55-9.

Segre, A. V., Consortium, D., investigators, M., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. 2010. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet,* 6.

Semple, R. K., Savage, D. B., Cochran, E. K., Gorden, P. & O'Rahilly, S. 2011. Genetic syndromes of severe insulin resistance. *Endocr Rev,* 32**,** 498-514.

Shen, W., Punyanitya, M., Wang, Z. M., Gallagher, D., St-Onge, M. P., Albu, J., Heymsfield, S. B. & Heshka, S. 2004. Visceral adipose tissue: relations between single-slice areas and total volume. *American Journal of Clinical Nutrition,* 80**,** 271-278.

Silander, K., Mohlke, K. L., Scott, L. J., Peck, E. C., Hollstein, P., Skol, A. D., Jackson, A. U., Deloukas, P., Hunt, S., Stavrides, G., *et al.* 2004. Genetic variation near the hepatocyte nuclear factor-4 alpha gene predicts susceptibility to type 2 diabetes. *Diabetes,* 53**,** 1141-9.

Skov, V., Glintborg, D., Knudsen, S., Jensen, T., Kruse, T. A., Tan, Q., Brusgaard, K., Beck-Nielsen, H. & Hojlund, K. 2007. Reduced expression of nuclear-encoded genes involved in mitochondrial oxidative metabolism in skeletal muscle of insulin-resistant women with polycystic ovary syndrome. *Diabetes,* 56**,** 2349-55.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., *et al.* 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature,* 445**,** 881-5.

Slatkin, M. 2009. Epigenetic inheritance and the missing heritability problem. *Genetics,* 182**,** 845-50.

Smits, P., Smeitink, J. & van den Heuvel, L. 2010. Mitochondrial translation and beyond: processes implicated in combined oxidative phosphorylation deficiencies. *J Biomed Biotechnol,* 2010**,** 737385.

Snijder, M. B., Visser, M., Dekker, J. M., Seidell, J. C., Fuerst, T., Tylavsky, F., Cauley, J., Lang, T., Nevitt, M. & Harris, T. B. 2002. The prediction of visceral fat by dual-energy X-ray absorptiometry in the elderly: a comparison with computed tomography and anthropometry. *International Journal of Obesity,* 26**,** 984-993.

Snogdal, L. S., Wod, M., Grarup, N., Vestmar, M., Sparso, T., Jorgensen, T., Lauritzen, T., Beck-Nielsen, H., Henriksen, J. E., Pedersen, O., *et al.* 2012. Common variation in oxidative phosphorylation genes is not a major cause of insulin resistance or type 2 diabetes. *Diabetologia,* 55**,** 340-8.

Spector, T. D. & Williams, F. M. 2006. The UK Adult Twin Registry (TwinsUK). *Twin Res Hum Genet,* 9**,** 899-906.

Spegel, P., Malmgren, S., Sharoyko, V. V., Newsholme, P., Koeck, T. & Mulder, H. 2011. Metabolomic analyses reveal profound differences in glycolytic and tricarboxylic acid cycle metabolism in glucose-responsive and -unresponsive clonal beta-cell lines. *Biochem J,* 435**,** 277-84.

Spencer, C. C., Su, Z., Donnelly, P. & Marchini, J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet,* 5**,** e1000477.

Steemers, F. J., Chang, W., Lee, G., Barker, D. L., Shen, R. & Gunderson, K. L. 2006. Whole-genome genotyping with the single-base extension assay. *Nat Methods,* 3**,** 31-3.

Stein, J. H., Korcarz, C. E., Hurst, R. T., Lonn, E., Kendall, C. B., Mohler, E. R., Najjar, S. S., Rembold, C. M. & Post, W. S. 2008. Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: A consensus statement from the American society of echocardiography carotid intima-media thickness task force endorsed by the society for vascular medicine. *Journal of the American Society of Echocardiography,* 21**,** 93-111.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., *et al.* 2007. Population genomics of human gene expression. *Nat Genet,* 39**,** 1217-24.

Strissel, K. J., Stancheva, Z., Miyoshi, H., Perfield, J. W., 2nd, DeFuria, J., Jick, Z., Greenberg, A. S. & Obin, M. S. 2007. Adipocyte death, adipose tissue remodeling, and obesity complications. *Diabetes,* 56**,** 2910-8.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

Suen, D. F., Norris, K. L. & Youle, R. J. 2008. Mitochondrial dynamics and apoptosis. *Genes Dev,* 22**,** 1577-90.

Szendroedi, J., Phielix, E. & Roden, M. 2012. The role of mitochondria in insulin resistance and type 2 diabetes mellitus. *Nat Rev Endocrinol,* 8**,** 92-103.

Tabak, A. G., Jokela, M., Akbaraly, T. N., Brunner, E. J., Kivimaki, M. & Witte, D. R. 2009. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *Lancet,* 373**,** 2215-21.

Tamaki, A., Ierano, C., Szakacs, G., Robey, R. W. & Bates, S. E. 2011. The controversial role of ABC transporters in clinical oncology. *Essays Biochem,* 50**,** 209-32.

Tanaka, T. 2005. [International HapMap project]. *Nihon Rinsho,* 63 Suppl 12**,** 29-34.

Tang, H., Liu, J., Niu, L., He, W. & Xu, Y. 2009. Variation in gene expression of presenilins-associated rhomboid-like protein and mitochondrial function in skeletal muscle of insulin-resistant rats. *Endocrine,* 36**,** 524-9.

Tapper, W. 2007. Linkage disequilibrium maps and location databases. *Methods Mol Biol,* 376**,** 23-45.

Tarasov, A. I., Nicolson, T. J., Riveline, J. P., Taneja, T. K., Baldwin, S. A., Baldwin, J. M., Charpentier, G., Gautier, J. F., Froguel, P., Vaxillaire, M., *et al.* 2008. A rare mutation in ABCC8/SUR1 leading to altered ATP-sensitive K+ channel activity and beta-cell glucose sensing is associated with type 2 diabetes in adults. *Diabetes,* 57**,** 1595-604.

Targher, G., Day, C. P. & Bonora, E. 2010. Risk of Cardiovascular Disease in Patients with Nonalcoholic Fatty Liver Disease. *New England Journal of Medicine,* 363**,** 1341-1350.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet,* 5**,** 123-35.

Toedling, J., Skylar, O., Krueger, T., Fischer, J. J., Sperling, S. & Huber, W. 2007. Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics,* 8**,** 221.

Toivonen, J. M., O'Dell, K. M., Petit, N., Irvine, S. C., Knight, G. K., Lehtonen, M., Longmuir, M., Luoto, K., Touraille, S., Wang, Z., *et al.* 2001. Technical knockout, a Drosophila model of mitochondrial deafness. *Genetics,* 159**,** 241-54.

Torkamani, A., Topol, E. J. & Schork, N. J. 2008. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics,* 92**,** 265-72.

Treuth, M. S., Hunter, G. R. & Kekesszabo, T. 1995. Estimating Intraabdominal Adipose-Tissue in Women by Dual-Energy X-Ray Absorptiometry. *American Journal of Clinical Nutrition,* 62**,** 527-532.

Triggs-Raine, B. L., Kirkpatrick, R. D., Kelly, S. L., Norquay, L. D., Cattini, P. A., Yamagata, K., Hanley, A. J., Zinman, B., Harris, S. B., Barrett, P. H., *et al.* 2002. HNF-1alpha G319S, a transactivation-deficient mutant, is associated with altered dynamics of diabetes onset in an Oji-Cree community. *Proc Natl Acad Sci U S A,* 99**,** 4614-9.

van den Ouweland, J. M., Lemkes, H. H., Ruitenbeek, W., Sandkuijl, L. A., de Vijlder, M. F., Struyvenberg, P. A., van de Kamp, J. J. & Maassen, J. A. 1992. Mutation in mitochondrial tRNA(Leu)(UUR) gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. *Nat Genet,* 1**,** 368-71.

Vaxillaire, M. & Froguel, P. 2008. Monogenic diabetes in the young, pharmacogenetics and relevance to multifactorial forms of type 2 diabetes. *Endocr Rev,* 29**,** 254-64.

Vionnet, N., Hani, E. H., Dupont, S., Gallina, S., Francke, S., Dotte, S., De Matos, F., Durand, E., Lepretre, F., Lecoeur, C., *et al.* 2000. Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. *Am J Hum Genet,* 67**,** 1470-80.

Virkamaki, A., Korsheninnikova, E., Seppala-Lindroos, A., Vehkavaara, S., Goto, T., Halavaara, J., Hakkinen, A. M. & Yki-Jarvinen, H. 2001. Intramyocellular lipid is associated with resistance to in vivo insulin actions on glucose uptake, antilipolysis, and early insulin signaling pathways in human skeletal muscle. *Diabetes,* 50**,** 2337-43.

Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. 2012. Five years of GWAS discovery. *Am J Hum Genet,* 90**,** 7-24.

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S., Thorleifsson, G., *et al.* 2010. Twelve

type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet,* 42**,** 579-89.

Vukkadapu, S. S., Belli, J. M., Ishii, K., Jegga, A. G., Hutton, J. J., Aronow, B. J. & Katz, J. D. 2005. Dynamic interaction between T cell-mediated beta-cell damage and beta-cell repair in the run up to autoimmune diabetes of the NOD mouse. *Physiol Genomics,* 21**,** 201-11.

Walder, K., Kerr-Bayles, L., Civitarese, A., Jowett, J., Curran, J., Elliott, K., Trevaskis, J., Bishara, N., Zimmet, P., Mandarino, L., *et al.* 2005. The mitochondrial rhomboid protease PSARL is a new candidate gene for type 2 diabetes. *Diabetologia,* 48**,** 459-68.

Wallace, M., Whelan, H. & Brennan, L. 2013. Metabolomic analysis of pancreatic beta cells following exposure to high glucose. *Biochim Biophys Acta,* 1830**,** 2583-90.

Wallace, T. M., Levy, J. C. & Matthews, D. R. 2004. Use and abuse of HOMA modeling. *Diabetes Care,* 27**,** 1487-95.

Wang, K., Li, M. & Bucan, M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet,* 81**,** 1278-83.

Wang, L., Jia, P., Wolfinger, R. D., Chen, X. & Zhao, Z. 2011. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics,* 98**,** 1-8.

Weir, G. C. & Bonner-Weir, S. 2004. Five stages of evolving beta-cell dysfunction during progression to diabetes. *Diabetes,* 53 Suppl 3**,** S16-21.

Wellcome Trust Case Control, C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

Winckler, W., Burtt, N. P., Holmkvist, J., Cervin, C., de Bakker, P. I., Sun, M., Almgren, P., Tuomi, T., Gaudet, D., Hudson, T. J., *et al.* 2005. Association of common variation in the HNF1alpha gene region with risk of type 2 diabetes. *Diabetes,* 54**,** 2336-42.

Winckler, W., Weedon, M. N., Graham, R. R., McCarroll, S. A., Purcell, S., Almgren, P., Tuomi, T., Gaudet, D., Bostrom, K. B., Walker, M., *et al.* 2007. Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes,* 56**,** 685-93.

Wittenhagen, L. M. & Kelley, S. O. 2002. Dimerization of a pathogenic human mitochondrial tRNA. *Nat Struct Biol,* 9**,** 586-90.

Wong, D. J., Nuyten, D. S., Regev, A., Lin, M., Adler, A. S., Segal, E., van de Vijver, M. J. & Chang, H. Y. 2008. Revealing targeted therapy for human cancer by gene module maps. *Cancer Res,* 68**,** 369-78.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-9.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., *et al.* 2011. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet,* 43**,** 519-25.

Yoshizumi, T., Nakamura, T., Yamane, M., Islam, A. H. M. W., Menju, M., Yamasaki, K., Arai, T., Kotani, K., Funahashi, T., Yamashita, S., *et al.* 1999. Abdominal fat: Standardized technique for measurement at CT. *Radiology,* 211**,** 283-286.

Yu, C., Chen, Y., Cline, G. W., Zhang, D., Zong, H., Wang, Y., Bergeron, R., Kim, J. K., Cushman, S. W., Cooney, G. J., *et al.* 2002. Mechanism by which fatty acids inhibit insulin activation of insulin receptor substrate-1 (IRS-1)-associated phosphatidylinositol 3-kinase activity in muscle. *J Biol Chem,* 277**,** 50230-6.

Yu, P., Ma, D. & Xu, M. 2005. Nested genes in the human genome. *Genomics,* 86**,** 414-22.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., *et al.* 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet,* 40**,** 638-45.

Zhang, W., Collins, A., Maniatis, N., Tapper, W. & Morton, N. E. 2002. Properties of linkage disequilibrium (LD) maps. *Proc Natl Acad Sci U S A,* 99**,** 17004-7.